NASA Technical Memorandum 101086

An Entropy-Based Approach to Nonlinear Stability

Marshal L. Merriam

March 1989

National Aeronautics and Space Administration

NASA Technical Memorandum 101086

An Entropy-Based Approach to Nonlinear Stability

Marshal L. Merriam, Ames Research Center, Moffett Field, California

March 1989

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure *Page*

AN ENTROPY-BASED APPROACH TO NONLINEAR STABILITY

Marshal L. Merriam

ABSTRACT

Many numerical methods used in Computational Fluid Dynamics (CFD) incorporate an artificial dissipation term to supress spurious oscillations and control nonlinear instabilities. The same effect can be accomplished by using upwind techniques, sometimes augmented with limiters to form Total Variation Diminishing (TVD) schemes. An analysis based on numerical satisfaction of the second law of thermodynamics allows many such methods to be compared and improved upon. For example, certain TVD schemes tend to "square" a smooth pulse. These can be detected a priori by their negative entropy production rates.

A nonlinear stability proof is given for discrete scalar equations arising from a conservation law. Solutions to such equations are bounded in the L_2 norm if the second law of thermodynamics is satisfied in a global sense over a periodic domain. It is conjectured that an analogous statement is true for discrete equations arising from systems of conservation laws.

Stability in the L_2 norm is not sufficient to exclude expansion shocks, oscillations, and other unphysical phenomena. Numerical experiments suggest that a more restrictive condition, a positive entropy production rate in each cell, is both necessary and sufficient to exclude such phenomena.

Construction of schemes which satisfy this condition is demonstrated for linear and nonlinear wave equations and for the one-dimensional Euler equations. For the linear wave equation all schemes which satisfy a cell entropy inequality are formally TVD. The converse is not true. The scheme for the Euler equations makes use of a remarkable identity to avoid the usual frozen coefficient approximation. Results are shown.

Since this form of dissipation is based on classical thermodynamics, it extends naturally to all types of fluid dynamics problems. In particular, artificial dissipation requirements for the Navier-Stokes equations, compressible and incompressible, may be quantitatively assessed.

Most existing forms of dissipation (including those mentioned above), treat problems in two and three dimensions as a sequence of one-dimensional operators. The present work does not require this approximation, and therefore extends more naturally to unstructured meshes. It holds promise for use with multigrid and grid refinement methods.

Chapter 1. INTRODUCTION

1.1 The Importance of Nonlinear Stability

For some time now it has been possible to obtain numerical solutions to physical problems described by systems of conservation laws. Among these problems is heat conduction, e.g., the determination of temperature distribution in a solid heated at a boundary. Existing numerical solution techniques for this problem are satisfactory in every respect. That is: they are fast, accurate, and robust. The same cannot be said of numerical solution techniques for problems in compressible fluid mechanics that are governed by the Navier-Stokes equations. While good comparison with experiment can sometimes be obtained, good error estimates usually cannot be.

Recent advances in Computational Fluid Dynamics (CFD) have made it possible to obtain reasonable solutions to the Navier-Stokes equations. Such solutions are generally sensitive to the details of the computational grid, to the time step, to the turbulence model, to the boundary conditions, and to the details of the artificial dissipation used. All of these are adjusted in practice to improve the agreement between the numerical solution and a corresponding experimental measurement or analytical result. The large number of adjustments and the lack of a good error measure make it hard to use CFD in cases where there are no experiments.

Navier-Stokes solvers tend to lack robustness. The more complicated the solver, the trickier it becomes to keep the computation stable and accurate. The usual scapegoats, cited with about equal frequency, are the grid, the time step, the turbulence model, the boundary conditions, and the artificial dissipation. In other words, selection of the proper adjustable parameters is critical. In the absence of theoretical guidance, a trial-and-error procedure is generally required to get a good answer.

Finally, there are questions of generality. There is considerable current interest in Navier-Stokes solutions involving chemically reacting flows about complex geometries. Algorithms which rely on a characteristic diagonalization are unsuitable when the characteristics are not known, real, or computable, as will surely be the case for chemically reacting flows. Similarly, algorithms which rely on dimensional splitting are difficult to apply to the unstructured grids being used to model complex geometries.

Numerical methods require speed, accuracy, and robustness to be useful. Faster computers improve the effective speed of a given method. Larger memories allow more grid points, thereby improving accuracy. Robustness, on the other hand, cannot be achieved through improvements in computer hardware. The aim of this work is to address the question of robustness. The approach explored here provides a physical, quantitative explanation for artificial viscosity. The improved understanding this requires provides insight

to other aspects of CFD. For example, the effects of time advance on stability are explored in Chapter 4 and a way of judging grid quality is introduced in Chapter 10.

1.2 Early Work in Computational Fluid Dynamics

The first attempts at numerical simulation of nonlinear convection were not completely successful. The solutions tended to develop numerical oscillations which grew exponentially as the solutions progressed in time. Approximating spatial derivatives proved to be especially troublesome; the most accurate formulae proved to be the least stable. What was lacking at the time was a physical and mathematical understanding of the convection process.

Much of this confusion was alleviated by an early monograph (ref. 1) in which Peter Lax detailed the theory of shock waves and sonic expansions. Among the ideas presented were the need for an entropy condition to make the solution unique, and the use of artificial viscosity to enforce an entropy condition. Domain of dependence ideas were introduced within the framework of characteristic theory. At the end of the work are some simple numerical methods and proofs that they lead to an accurate answer (in the limit of very fine grids). The ideas presented in Lax's monograph have been very important in the development of CFD. Their influence is apparent in every major scheme in use today, and in the work presented here.

An early success was the work of Murman and Cole (ref. 2) on the transonic small disturbance equations. The main advance in this work was the Murman-Cole switch which used domain of dependence ideas, cell-by-cell, to choose between two approximations for the streamwise derivatives. One stencil would be used in subsonic flow, another in supersonic flow. Though limited to transonic or subsonic flow about slender bodies with attached flow, this work represented a tremendous advance over the calculations that preceded it. To quote Robert MacCormack (ref. 3) " The use of these methods spread rapidly and airfoil sections were no longer designed experimentally by cutting and filing."

In the mid-seventies, the first solutions to the Navier-Stokes equations appeared. Central to this advance was the work of Steger (ref. 4) and Pulliam and Steger (ref. 5) in their ARC2D and ARC3D codes. In addition to the use of generalized coordinates (but not generalized topologies), these codes used implicit time advance and an artificial dissipation based on fourth differences. They also used the Baldwin-Lomax turbulence model (ref. 6), a simple, empirical model that replaced turbulence with an equivalent eddy viscosity. Although suffering from oscillations near shocks and having very poor convergence by today's standards, these codes represented a revolutionary advance in the ability to calculate separated flows and model high Reynolds number flows.

In the late seventies and early eighties, algorithms began to appear (refs. 7,8, and 9) which incorporated a characteristic decomposition of the inviscid fluxes. It was noted

that the flux Jacobian with respect to the conservative variables had real eigenvalues and eigenvectors which were not too difficult to compute. This allowed the use of domain of dependence ideas in the Euler equations. The algorithms which resulted had no need for artificial dissipation of the type used in ARC2D as they were naturally dissipative. This feature removed the need to choose the coefficient for artificial dissipation. It also removed any control over it. In spite of the major increase in physical understanding included in these algorithms, they continued to exhibit minor numerical oscillations which had no physical basis.

To get rid of these annoying oscillations, researchers attacked them head on, producing schemes which, at least for scalar equations in one dimension, were guaranteed not to have any wiggles. The first scheme of this type, Flux Corrected Transport (ref. 10) demonstrated the concept. Later the concept of Total Variation Diminishing (TVD) schemes was formalized by Harten (ref. 11) and implemented for gasdynamics. Although Lax referred to proofs (refs. 12 and 13) that the desired physical solution has the TVD property, it does not follow that all TVD schemes converge to a physical solution. Indeed, schemes have been introduced (ref. 2) which satisfy the TVD condition and yet support unphysical expansion shocks. In the rush to kill off those annoying wiggles, some researchers have ignored the entropy inequality that is fundamental to getting a correct solution.

In this decade, some research has been done on schemes which satisfy an entropy inequality. Some of this work, done by Osher (ref. 14) and Tadmor (ref. 15), will be explored in later chapters. It concerns a numerical approximation to the entropy inequality on a cell-by-cell basis. From a different perspective, Hughes, Franca, and Mallet (ref. 16) have concentrated on a finite element method which guarantees global satisfaction of the second law of thermodynamics in the steady state. The local oscillations which still occur offer evidence of local violations of the second law, even though it is globally satisfied. The thrust of the work presented here will be to satisfy the second law of thermodynamics on a cell-by-cell basis.

1.3 Thesis Organization and Summary of Results

The goal of this research is to analyze the requirements for nonlinear stability and produce an algorithm to show that these can be met. Chapter 2 establishes notation and shows some important properties of partial differential equations (PDE's) which come from conservation laws. These properties are exploited to show that solutions of scalar hyperbolic equations in multidimensions are bounded for all time. For systems of equations, analysis shows only that the total entropy in the domain is bounded; the question of bounds on the solution remains open.

In Chapter 3 the nonlinear stability property of the PDE is extended to the semi-discrete difference equations. The semi-discrete analog of the second law is derived. Chap-

ter 3 goes on to discuss, in general terms, how to construct space differencing schemes which satisfy a cell entropy inequality. Such schemes can be constructed with at least second order accuracy.

In Chapter 4 the nonlinear stability property of the PDE is extended to the fully discrete difference equations. It is shown that satisfying a semi-discrete cell entropy inequality is sufficient to satisfy a fully discrete one; provided that implicit Euler time advance is used. A modified Crank-Nicolson time advance for scalar equations also has this property.

In Chapters 5 through 7, examples are given for problems that increase in complexity up to the quasi-one-dimensional Euler equations. These examples show how the ideas given in Chapters 2, 3, and 4 can be applied in practice. Chapters 8 and 9 show how these new ideas relate to existing schemes.

In Chapter 10 a good deal of future work is outlined. A scheme is not very useful if it has limited applicability. On the other hand, it is not worthwhile to extend an impractical scheme; better to clean up the simple case first. For this reason, Chapter 10 is limited to describing in some detail how these extensions might be made and discussing the implications of these ideas on several of the major problems that still confront us in CFD. The author feels that these ideas constitute a powerful new tool, making possible a whole new approach to these problems. Some obvious approaches will be outlined. Finally, Chapter 11 provides a summary.

Chapter 2. STABILITY AND THE SECOND LAW

2.1 A Physical Description of Entropy

The Euler equations for gasdynamics in one dimension have a well-defined conservation form and do not allow heat transfer. They also have no viscous terms. In one dimension the Euler equations can be written

$$q_{,t} + f_{,x} = 0 \tag{2.1.1}$$

where

$$q = \begin{bmatrix} \rho \\ \rho u \\ e \end{bmatrix} \quad \text{and} \quad f = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ (e+p)u \end{bmatrix} \tag{2.1.2}$$

The quantities ρ and u represent the local mass density and the local velocity respectively. The quantity e represents the total energy density. The total energy can be expressed as the sum of the thermal and kinetic energies of the gas.

$$e = \rho c_v T + \frac{1}{2}\rho u^2 \tag{2.1.3}$$

The variable T represents the local temperature and c_v represents the heat capacity at constant volume. For a thermally and calorically perfect gas, $p = \rho RT$. This equation of state, and the relations $\gamma = \frac{c_p}{c_v}$ and $R = c_p - c_v$, allow derivation of an expression for pressure.

$$p = (\gamma - 1)(e - \frac{1}{2}\rho u^2) \tag{2.1.4}$$

Equations (2.1.1) are formally correct only as long as the derivatives are defined. The corresponding integral equations are more lenient. These are

$$\int_V q[t + \Delta t]\, dv - \int_V q[t]\, dv + \int_t^{t+\Delta t} \oint_{\partial V} \mathbf{f} \cdot \mathbf{n}\, dA\, d\tau = 0 \tag{2.1.5}$$

where $\mathbf{n}$ represents the outward facing unit normal to the surface ∂V (in one dimension the surface integral reduces to a difference). To derive (2.1.1) from (2.1.5), the divergence theorem is used to convert the surface integrals into volume integrals. The limit is then taken as the space time control volume becomes infinitesimally small. At this limit, the integrand must vanish if the integral vanishes. Such a step is valid only in the absence of a discontinuity. Because of this caveat, the use of (2.1.1) is prone to difficulties when applied to flows with discontinuities.

Equation (2.1.5) represents three separate conservation laws, one for each component of q. The first equation corresponds to conservation of mass, the second corresponds

to conservation of momentum, and the third equation is a statement of the first law of thermodynamics.

$$\int_V e[t+\Delta t]\,dv - \int_V e[t]\,dv + \int_t^{t+\Delta t} \oint_{\partial V} (e+p)\mathbf{u}\cdot\mathbf{n}\,dA\,d\tau = 0 \qquad (2.1.6)$$

The conserved quantity in (2.1.6) is the total energy, the sum of the internal energy density ($\rho c_v T$) and the kinetic energy density ($\frac{1}{2}\rho u^2$) integrated over any control volume. These two energy forms are not (and should not be) individually conserved; the first law allows arbitrary transfers of energy between them.

A fluid which is solely governed by the conservation laws in (2.1.5) can display very unusual behavior. For example, it can transfer all of its internal energy to kinetic energy. The result is a very fast, very cold jet of air. This Carnot air conditioner is allowed under conservation of mass, momentum and energy, yet is never observed in practice. To explain this discrepancy requires introducing the second law of thermodynamics. Under the second law, certain transfers between internal energy and kinetic energy are not possible.

Three independent variables are required to describe the fluid of (2.1.5). A fundamental result of thermodynamics suggests that such a gas has two reversible work modes (ref. 17). These are known for the one-dimensional Euler equations. If work is expended to increase fluid kinetic energy at constant pressure and temperature, it can be completely recovered by decelerating the flow. This is called an isentropic acceleration and deceleration. If work is expended to compress the fluid isentropically at constant velocity (raising its temperature in the process), it can be completely recovered in a subsequent expansion (which lowers the temperature). Through these processes, the kinetic energy and internal energy (temperature) can be independently and reversibly varied. Other reversible state changes are a combination of these two processes. For example, the work required for an isentropic compression can be provided by an isentropic deceleration. Similarly, the work required for an isentropic acceleration can be provided by an isentropic expansion.

In addition to reversible state changes, there are irreversible ones. For example, heat transfer across a finite temperature difference is irreversible as is viscous momentum transfer. Although not explicitly included in (2.1.5), these mechanisms are significant within the profile of shocks, the main source of irreversibility in the Euler equations. The word irreversible implies that certain state changes (like transferring heat against a temperature gradient) are not possible; they are never observed in the laboratory on macroscopic scales. The Carnot air conditioner described above requires the use of state changes that are impossible in this sense.

The second law of thermodynamics provides a way of discriminating between the possible state changes and the impossible. It can be written in a form similar to (2.1.5)

$$\int_t^{t+\Delta t} \int_V \dot{P}_s \, dv \, d\tau \equiv \int_V S[t+\Delta t] \, dv - \int_V S[t] \, dv + \int_t^{t+\Delta t} \oint_{\partial V} \mathbf{F} \cdot \mathbf{n} \, dA \, d\tau \geq 0 \quad (2.1.7)$$

In this equation, $\dot{P}_s$ represents the rate of production of entropy per unit volume, S is the entropy per unit volume, and $\mathbf{F}$ is the entropy flux per unit area. For reversible processes, $\dot{P}_s = 0$. For irreversible processes $\dot{P}_s > 0$. Processes for which $\dot{P}_s < 0$ are never observed.

The quantities S and F depend on the fluid being studied and on the idealizations being applied. For the Euler equations in the absence of heat transfer, these are

$$S = \rho s \quad (2.1.8)$$

$$\mathbf{F} = \rho \mathbf{u} s \quad (2.1.9)$$

where s is the nondimensional specific entropy.

From Reynolds (ref. 18), comes the expression for the entropy difference between an arbitrary state, q' and a reference state q_0, for a perfect gas.

$$\frac{s' - s_0}{c_v} = \log\left(\frac{p'/p_0}{(\rho'/\rho_0)^\gamma}\right) \quad (2.1.10)$$

Equation (2.1.10) suggests nondimensional forms for entropy, density and pressure. Using these gives

$$s = \log\left(\frac{p}{\rho^\gamma}\right) \quad (2.1.11)$$

which is used throughout this work. The second law (2.1.7) is not limited to the Euler equations or to an ideal gas. It applies to all of continuum mechanics.

2.2 A Mathematical Description of Entropy

In addition to the physical view of entropy given above, there is a well-developed mathematical theory of entropy (from Lax (ref. 19) and Krushkov (ref. 20)) given in the monograph previously mentioned (ref. 1). This theory, now well known, holds that any entropy pair (S, F) must have two properties.

$$S_{,qq} < 0 \quad (2.2.1)$$

$$F_{,q} = S_{,q} f_{,q} \quad (2.2.2)$$

where S is the the entropy per unit volume and F is the entropy flux.

The quantities which appear in (2.2.1) and (2.2.2) are defined as follows

$$S_{,qq} \equiv \begin{bmatrix} \frac{\partial^2 S}{\partial q_1 \partial q_1} & \frac{\partial^2 S}{\partial q_1 \partial q_2} & \cdot & \frac{\partial^2 S}{\partial q_1 \partial q_n} \\ \frac{\partial^2 S}{\partial q_2 \partial q_1} & \frac{\partial^2 S}{\partial q_2 \partial q_2} & \cdot & \frac{\partial^2 S}{\partial q_2 \partial q_n} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{\partial^2 S}{\partial q_n \partial q_1} & \frac{\partial^2 S}{\partial q_n \partial q_2} & \cdot & \frac{\partial^2 S}{\partial q_n \partial q_n} \end{bmatrix} \quad \text{and} \quad f_{,q} = \begin{bmatrix} \frac{\partial f_1}{\partial q_1} & \frac{\partial f_1}{\partial q_2} & \cdot & \frac{\partial f_1}{\partial q_n} \\ \frac{\partial f_2}{\partial q_1} & \frac{\partial f_2}{\partial q_2} & \cdot & \frac{\partial f_2}{\partial q_n} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{\partial f_n}{\partial q_1} & \frac{\partial f_n}{\partial q_2} & \cdot & \frac{\partial f_n}{\partial q_n} \end{bmatrix} \tag{2.2.3}$$

$$S_{,q} \equiv \left[\frac{\partial S}{\partial q_1}, \quad \frac{\partial S}{\partial q_2}, \quad \cdots, \quad \frac{\partial S}{\partial q_n} \right] \quad \text{and} \quad F_{,q} \equiv \left[\frac{\partial F}{\partial q_1}, \quad \frac{\partial F}{\partial q_2}, \quad \cdots, \quad \frac{\partial F}{\partial q_n} \right] \tag{2.2.4}$$

The convexity condition (2.2.1) forces irreversible processes to run in the correct direction and produce entropy. A consequence of the compatability condition (2.2.2) is that reversible processes do not produce entropy. The following two examples illustrate these effects for the Euler equations.

It is well known that if a quantity of fluid is placed in an isolated box, the equilibrium condition will be the one for which entropy is a maximum. This occurs when all of the fluid is at the same state. This state, $\overline{q}$, is simply the constant state with the correct total mass, momentum and energy for the box.

$$\overline{q} \equiv \frac{1}{V} \int_V q \, dv \tag{2.2.5}$$

It is the nature of isolated boxes that there is no flux through their boundaries. This, combined with the conservation law (2.1.5), is sufficient to guarantee that $\overline{q}$ doesn't change with time. The state transition from $q(\mathbf{x})$ to $\overline{q}$ is physically irreversible so the entropy contained within the volume should increase. This property follows mathematically if the convexity condition is met, as can be shown by using Taylor's theorem with remainder as follows.

$$S = \bar{S} + \bar{S}_{,q}(q - \overline{q}) + \frac{1}{2}(q - \overline{q})^T S^{**}_{,qq}(q - \overline{q}) \tag{2.2.6}$$

Here the quantity $\overline{q}$ is used to define $\bar{S} \equiv S(\overline{q})$ and $\bar{S}_{,q} \equiv S_{,q}(\overline{q})$. The term $S^{**}_{,qq}$ is evaluated at some unspecified state, q^{**}, which makes (2.2.6) an equality. The existence of such a state is guaranteed by the intermediate value theorem.

Both sides of (2.2.6) can be integrated over a control volume (which encloses the box) to give

$$\int_V S \, dv = \int_V \bar{S} \, dv + \bar{S}_{,q} \int_V (q - \overline{q}) \, dv + \frac{1}{2} \int_V (q - \overline{q})^T S^{**}_{,qq}(q - \overline{q}) \, dv \tag{2.2.7}$$

The second term on the right-hand side vanishes by the definition of $\bar{q}$.

$$\int_V S\,dv = \int_V \bar{S}\,dv + \frac{1}{2}\int_V (q-\bar{q})^T S^{**}_{,qq}(q-\bar{q})\,dv \tag{2.2.8}$$

The integrand of the last term is a proper quadratic form since $S^{**}_{,qq}$ is a negative definite matrix. This implies that the last term is nonpositive. In this way, convexity is used to show that the maximum entropy occurs when the equality holds, i.e., when $q = \bar{q}$ over the entire cell.

The compatability condition (2.2.2) addresses reversible processes. The Euler equations (2.1.5) do not contain terms for heat transfer, viscosity or any other dissipation mechanism. As long as the solution is smooth, the flow processes are reversible and no entropy production should occur. Such a statement follows mathematically if the compatability condition is met. This is shown by first multiplying (2.1.1) on the left by $S_{,q}$.

$$S_{,q}q_{,t} + S_{,q}f_{,x} = 0 \tag{2.2.9}$$

In smooth regions (where the derivatives are finite) the chain rule can be applied as follows

$$S_{,t} + S_{,q}f_{,q}q_{,x} = 0 \tag{2.2.10}$$

At this point we assume that (2.2.2) holds and make the substitution.

$$S_{,t} + F_{,q}q_{,x} = 0 \tag{2.2.11}$$

Using the chain rule one more time leads to

$$S_{,t} + F_{,x} = 0 \tag{2.2.12}$$

which is the differential statement of conservation of entropy. This shows that if (2.2.2) holds and if the solution is smooth, then entropy will be conserved and $\dot{P}_s = 0$. Such a statement is known to be true for the Euler equations (ref. 21), in which entropy production is nonzero only at discontinuities.

The definitions of S and F are not always unique. For example, Harten (ref. 22) showed a family of entropy pairs for the Euler equations. Furthermore (again from Lax (ref. 1)) if q contains more than two independent variables, (2.2.2) represents a system of partial differential equations that is overdetermined and may have no solution. Fortunately, for the specific case of the Navier-Stokes equations in three dimensions, Mallet (ref. 23) showed that S and F exist and are essentially unique.

2.3 A Relationship Between Mathematical and Physical Entropy

At this point, a question arises: Does the mathematical entropy described in Section 2.2 correspond to the physical entropy described in Section 2.1? At least for some important cases, the answer is yes. For the one-dimensional Euler equations, q and f are defined as in (2.1.2)

$$q = \begin{bmatrix} \rho \\ \rho u \\ e \end{bmatrix} \qquad f = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ (e+p)u \end{bmatrix} \tag{2.3.1}$$

As mentioned in Section 2.1, the entropy pair (S, F) suggested by thermodynamics is

$$S = \rho s \tag{2.3.2}$$

$$F = \rho u s \tag{2.3.3}$$

where

$$s = \log\left(\frac{p}{\rho^\gamma}\right) \tag{2.3.4}$$

Given these definitions for q, f, S, and F, it is possible to check the physical entropy and entropy flux against the mathematical constraints (2.2.1) and (2.2.2). The first check is whether $S_{,qq} < 0$. The derivatives are easily carried out using the symbolic algebra program MACSYMA; the matrix is

$$S_{,qq} = -\left(\frac{(\gamma-1)}{p}\right)^2 \begin{bmatrix} \frac{\rho u^4}{4} + \frac{\gamma p^2}{(\gamma-1)^2\rho} & -\frac{\rho u^3}{2} & \rho u^2 - e \\ -\frac{\rho u^3}{2} & \frac{\rho u^2}{2} + e & -\rho u \\ \rho u^2 - e & -\rho u & \rho \end{bmatrix} \tag{2.3.5}$$

The determinant of $S_{,qq}$ is also easily computed by MACSYMA and is

$$|S_{,qq}| = -\frac{(\gamma-1)^4}{p^3} \tag{2.3.6}$$

which indicates that $S_{,qq}$ is nonsingular, at least for $p > 0$. Furthermore there exists, for the Euler equations, a matrix Y^{-1} such that

$$S_{,qq} = -(Y^{-1})^T(Y^{-1}) \tag{2.3.7}$$

This implies that $S_{,qq}$ is negative semi-definite. Since (2.3.6) rules out the singular case, $S_{,qq}$ is, in fact, negative definite, demonstrating that (2.2.1) is satisfied when S is given by (2.3.2).

The matrices Y and Y^{-1} are given below. For convenience, the speed of sound $c = \sqrt{\frac{\gamma p}{\rho}}$ has been introduced.

$$Y = \begin{bmatrix} \beta_1 & \beta_2 & \beta_2 \\ \beta_1 u & \beta_2(u-c) & \beta_2(u+c) \\ \frac{\beta_1 u^2}{2} & \beta_2(\frac{u^2}{2} - uc + \frac{c^2}{(\gamma-1)}) & \beta_2(\frac{u^2}{2} + uc + \frac{c^2}{(\gamma-1)}) \end{bmatrix} \tag{2.3.8}$$

$$Y^{-1} = \left(\frac{1}{2c^2}\right) \begin{bmatrix} \frac{2c^2-(\gamma-1)u^2}{\beta_1} & \frac{2(\gamma-1)u}{\beta_1} & \frac{-2(\gamma-1)}{\beta_1} \\ \frac{(\gamma-1)u^2}{2\beta_2} + \frac{uc}{\beta_2} & \frac{-(\gamma-1)u-c}{\beta_2} & \frac{(\gamma-1)}{\beta_2} \\ \frac{(\gamma-1)u^2}{2\beta_2} - \frac{uc}{\beta_2} & \frac{-(\gamma-1)u+c}{\beta_2} & \frac{(\gamma-1)}{\beta_2} \end{bmatrix} \tag{2.3.9}$$

$$\beta_1 = \sqrt{\frac{\rho}{\gamma}}, \quad \beta_2 = \sqrt{\frac{\rho}{2\gamma(\gamma-1)}} \tag{2.3.10}$$

Identity (2.3.7) has been verified by construction using MACSYMA. This identity is a new result. It is discussed further in Chapter 7.

The second requirement on the mathematical entropy is the existence of an entropy flux F such that

$$F_{,q} = S_{,q} f_{,q} \tag{2.3.11}$$

Such an entropy flux is given by (2.3.3). This is easily verified with MACSYMA. For reference the individual terms are given here.

$$F_{,q} = \left[\frac{(\gamma-1)\rho u^3}{2p} - \gamma u, \quad s - \frac{(\gamma-1)\rho u^2}{p}, \quad \frac{(\gamma-1)\rho u}{p} \right] \tag{2.3.12}$$

$$S_{,q} = \left[s - (\gamma+1) + \frac{(\gamma-1)e}{p}, \quad -\frac{(\gamma-1)\rho u}{p}, \quad \frac{(\gamma-1)\rho}{p} \right] \tag{2.3.13}$$

$$f_{,q} = \begin{bmatrix} 0 & 1 & 0 \\ \frac{(\gamma-3)u^2}{2} & -(\gamma-3)u & (\gamma-1) \\ (\gamma-1)u^3 - \frac{\gamma e u}{\rho} & \frac{\gamma e}{\rho} - \frac{3}{2}(\gamma-1)u^2 & \gamma u \end{bmatrix} \tag{2.3.14}$$

Since satisfaction of both requirements has been verified, (2.3.2) and (2.3.3) define a valid entropy pair for the Euler equations in the mathematical sense. It is well known that this entropy pair also describes the physical entropy density and entropy flux (ref. 23).

This example, and those of the previous section provide indications that the mathematical and physical entropy statements are compatible; if one is true then the other is true. Such a conclusion is not new; these examples are included for the benefit of readers unfamiliar with this somewhat esoteric subject.

2.4 The Sign Convention for Entropy

When entropy was first explored from a physical point of view by Carnot and others, the entropy statement was viewed almost theologically. The entropy of the universe is increasing; everything goes toward chaos. Given this point of view, it seems natural that entropy production should be a positive quantity. For scalar conservation laws with a single dependent variable u, $S = -u^2$. The second law requires that this increase with time for any closed system.

By contrast, the first mathematical treatments of conservation laws were concerned with bounds on the solution. The natural statement of such bounds for scalar conservation laws is that u^2 is conserved or decreases with time. This seemed more natural than an equivalent statement about $-u^2$. When the concept of entropy was later introduced, consistency of notation demanded that entropy should decrease with time. As a consequence, many papers on the subject consider mathematical entropy to be a decreasing quantity, positive entropy production rates to be impossible and $S_{,qq}$ to be positive definite.

This difference in sign convention is pointed out to help the interested reader avoid confusion when reviewing the literature on this subject. There is no substantive difference between the two points of view. As a purely arbitrary choice, this work adheres to the physical convention.

2.5 Nonlinear Stability

Nonlinear stability is the holy grail of numerical analysis. Always sought, occasionally sighted, this goal has eluded researchers for many years. In this study, a promising approach is examined in some detail. Some of the ideas in this approach are due to Dutt (ref. 24) who several years ago offered a stability proof subject to certain questionable assumptions which will be pointed out along the way. What follows is a greatly simplified and shortened version of his proof, following the same general outline. Though a complete proof still eludes us, the goal seems much closer than it was.

The basic idea is to first show that the total entropy in the domain is bounded. This can be done by showing that it steadily increases and that it cannot increase above a certain level. The next step is to show that if the entropy is bounded, the solution is bounded in a certain norm.

CONJECTURE.

Over some domain which encloses volume Ω and is enclosed by a surface $\partial\Omega$

Given:

$$\oint_{\partial\Omega} \mathbf{f} \cdot \mathbf{n}\, dA = 0 \tag{2.5.1}$$

$$\int_\Omega q[t]\,dV - \int_\Omega q[0]\,dV + \int_0^t \oint_{\partial\Omega} \mathbf{f}\cdot\mathbf{n}\,dA\,d\tau = 0 \tag{2.5.2}$$

$$\oint_{\partial\Omega} \mathbf{F}\cdot\mathbf{n}\,dA = 0 \tag{2.5.3}$$

$$\int_\Omega S[t]\,dV - \int_\Omega S[0]\,dV + \int_0^t \oint_{\partial\Omega} \mathbf{F}\cdot\mathbf{n}\,dA\,d\tau \geq 0 \tag{2.5.4}$$

and defining

$$\overline{q}[t] \equiv \frac{1}{\Omega}\int_\Omega q[t]\,dV \tag{2.5.5}$$

There exists a value q^* which can be determined from the initial conditions $q[0]$ such that

$$0 < -\frac{1}{2}\int_\Omega (q-\overline{q})^T S^*_{,qq}(q-\overline{q})\,dV \leq \int_\Omega S(\overline{q}[0]) - S(q[t])\,dV \leq \int_\Omega S(\overline{q}[0]) - S(q[0])\,dV \tag{2.5.6}$$

Notice that $-\frac{1}{2}\int_\Omega (q-\overline{q})^T S^*_{,qq}(q-\overline{q})\,dV$ is a norm, since $S^*_{,qq}$ is negative definite. It is in this norm that the solution remains bounded.

REASONING:

Combining the conservation law (2.5.2) with the periodicity constraint (2.5.1) gives the relation

$$\int_\Omega q[t]\,dV = \int_\Omega q[0]\,dV \tag{2.5.7}$$

which says that the total amount of q in the domain is a constant over time. Dividing by the domain volume and applying definition (2.5.5) to both sides gives

$$\overline{q}[t] = \overline{q}[0] \tag{2.5.8}$$

which is to say that $\overline{q}$ is also a constant over time. This being the case, the argument can be dropped and $\overline{q}[t]$ can be expressed simply as $\overline{q}$.

At this point an earlier result (2.2.8) can be invoked using the entire domain as the control volume.

$$\int_\Omega S[t]\,dV = \int_\Omega \bar{S}\,dV + \frac{1}{2}\int_\Omega (q-\overline{q})^T S^{**}_{,qq}(q-\overline{q})\,dV \tag{2.5.9}$$

As before $\bar{S} \equiv S(\overline{q})$. Also, the quadratic term is negative at each point in the domain so

$$\int_\Omega \bar{S}\,dV \geq \int_\Omega S[t]\,dV \tag{2.5.10}$$

Combining (2.5.4) with (2.5.3) implies that

$$\int_\Omega S[t]\,dV \geq \int_\Omega S[0]\,dV \tag{2.5.11}$$

which implies that the total amount of entropy in the domain is a nondecreasing function of time. Combining (2.5.11) with (2.5.10) gives

$$\int_\Omega \bar{S}\,dV \geq \int_\Omega S[t]\,dV \geq \int_\Omega S[0]\,dV \tag{2.5.12}$$

so the total entropy in the domain is bounded for all time. Note that both bounds can be computed from the initial conditions. Subtracting the upper bound from each term and multiplying by -1 gives

$$\int_\Omega (\bar{S} - S[0])\,dV \geq \int_\Omega (\bar{S} - S[t])\,dV \geq 0 \tag{2.5.13}$$

In Dutt's proof, this is an intermediate result. There is a significant difference, however, in that he uses a particular value of $\bar{q}$ which does not depend solely on the initial conditions. Rather, he limits his proof to the Navier-Stokes equations and assumes bounds on the density and temperature for all time. This results in a less precise bound and, more importantly, assumes a good deal of what he is trying to prove.

This concludes the first portion of the proof. Equation (2.5.13) guarantees a bound on $\int_\Omega (\bar{S} - S[t])\,dV$. The second portion of the proof deals with the stability implications of this statement.

Since the first portion of the proof dealt with domain integrated quantities, a suitable stability definition is

$$||q(t)|| \leq \kappa ||q(0)|| \tag{2.5.14}$$

for $0 \leq t \leq T$. Any norm can be used, including the L_2 norm. It is assumed that the scheme integrates up to a fixed time $T > t_0$. It would be nice if κ was near unity instead of some large value, but this is not required.

The middle term of (2.5.13) can be calculated using (2.5.9).

$$\int_\Omega (\bar{S} - S[0])\,dV \geq \int_\Omega (\bar{S} - S[t])\,dV = -\frac{1}{2}\int_\Omega (q-\bar{q})^T S^{**}_{,qq}(q-\bar{q})\,dV \geq 0 \tag{2.5.15}$$

Notice that $-\frac{1}{2}\int_\Omega (q-\bar{q})^T S^{**}_{,qq}(q-\bar{q})\,dV$ is almost a suitable norm. It is always positive, since the integrand is a proper quadratic form and $S^{**}_{,qq}$ is negative definite. The problem is that q^{**} is not determined. Furthermore it is different for each control volume.

These problems can be overcome if a single q^* exists such that

$$-\frac{1}{2}\int_\Omega (q-\overline{q})^T S^{**}_{,qq}(q-\overline{q})\,dV \geq -\frac{1}{2}\int_\Omega (q-\overline{q})^T S^{*}_{,qq}(q-\overline{q})\,dV \geq 0 \tag{2.5.16}$$

where $S^*_{,qq} \equiv S_{,qq}(q^*)$. The state q^* must not be a function of time. Therefore it must depend only on the initial conditions. If such a state can be found, then a suitable norm is

$$\|q\|^2 = -\frac{1}{2}\int_\Omega (q-\overline{q})^T S^{*}_{,qq}(q-\overline{q})\,dV \tag{2.5.17}$$

Such a norm is bounded for all time by the relation

$$\int_\Omega (\bar{S}-S[0])\,dV \geq \int_\Omega (\bar{S}-S[t])\,dV = -\frac{1}{2}\int_\Omega (q-\overline{q})^T S^{**}_{,qq}(q-\overline{q})\,dV \geq \|q\|^2 \geq 0 \tag{2.5.18}$$

It is not obvious how to find q^* in general, which is why this is a conjecture and not a theorem. Some comments are in order, however. First, (2.2.6) is exact. A similar statement which is only approximately true is

$$S(q) \approx \bar{S} - \bar{S}_{,q}(q-\overline{q}) + \frac{1}{2}(q-\overline{q})^T \bar{S}_{,qq}(q-\overline{q}) \tag{2.5.19}$$

This comes from using a truncated Taylor series instead of Taylor's theorem with remainder so $S^{**}_{,qq} \approx \bar{S}_{,qq}$, with the approximation improving steadily as $q[t]$ approaches $\overline{q}$. This suggests that $\overline{q}$ might meet the requirements for q^*. It can be determined from the initial conditions and

$$-\frac{1}{2}\int_\Omega (q-\overline{q})^T \bar{S}_{,qq}(q-\overline{q})\,dV \geq 0 \tag{2.5.20}$$

The remaining requirement, (2.5.16), may be true, but is not easily proved in general.

A second comment is that $-S^{**}_{,qq}$ is positive definite. The evaluation state q^{**} varies over the time and space domains of the problem. Presumably there is then some minimum value of $S^{**}_{,qq}$. (This is tacitly assumed by Dutt in his proof.) The value of q at which this occurs could then be used for q^*. Such a value of q^* meets the requirements of (2.5.16) but is not a function of the initial conditions. Such an a posteriori norm has no predictive value. This is because it may assign a very small weight to a mode which is growing exponentially while assigning larger weights to modes which decay. While such a norm may decay for a time, it will eventually grow without bound.

In summary then, this work differs from Dutt's in everything but the general outline. It removes some unnecessary assumptions and an unjustified one. It replaces general bounds with specific ones. Most important, the analysis is more generally applicable and easier to follow than the pioneering work of Dutt. An unfortunate effect of all this progress is the downgrading of the result from a proof to a conjecture, a consequence of removing the assumption that q^* exists in general.

Scalar Equations

There is considerable work in the literature concerning scalar PDEs of the form

$$u_{,t} + f_{,x} = 0 \tag{2.5.21}$$

where u is traditionally used in place of q in this context. It is easy to construct a valid entropy pair for such equations. The entropy

$$S(u) \equiv -u^2 \tag{2.5.22}$$

This certainly meets the convexity requirement since

$$S_{,uu} = -2 < 0 \tag{2.5.23}$$

Finally the compatability constraint

$$F_{,u} = S_{,u} f_{,u} \tag{2.5.24}$$

can be used to define F, if $f(u)$ is differentiable. The quantities $S_{,u}$ and $f_{,u}$ can be constructed, multiplied, and integrated to give $F(u)$.

The importance of this rather obvious result is that the above conjecture becomes a theorem for scalar equations. This follows because $S_{,uu}$ is a constant, independent of the evaluation point. The result is that for scalars

$$\int_\Omega (\bar{S} - S[0])\, dV \geq \|(q - \bar{q})\|_2^2 \geq 0 \tag{2.5.25}$$

In fact, using (2.5.22) and (2.5.12) gives

$$\|\bar{q}\|_2^2 \geq \|q(t)\|_2^2 \geq \|q(0)\|_2^2 \tag{2.5.26}$$

Using the L_2 norm and $\kappa = 1$, (2.5.14) holds. Thus, the second law is sufficient for L_2 stability of all scalar hyperbolic PDEs. This result holds for multidimensional periodic domains and is probably extendable to unbounded domains as well.

Euler Equations

In the case of the Euler equations in one dimension, the conjecture has an interesting interpretation. Recall from Section 2.3 that for these equations

$$S_{,qq} = -(Y^{-1})^T (Y^{-1}) \tag{2.5.27}$$

If the conjecture is true, q^* exists and one can construct new variables of the form

$$\tilde{q} \equiv Y^{-1} q \tag{2.5.28}$$

where the matrix Y^{-1} is always evaluated at q^*. The average values, $\overline{q}$, can also be subjected to this transformation. If q^* exists the result can be written

$$\int_\Omega (\bar{S} - S[0])\, dV \geq \frac{1}{2} \|(\widetilde{q - \overline{q}})\|_2^2 \geq 0 \tag{2.5.29}$$

Thus, the second law is sufficient for L_2 stability of the one-dimensional Euler equations in the transformed variables, provided that the conjecture holds.

The author has made some effort to establish the existence of q^* in this case. In particular, an attempt was made to show that

$$S^*_{,qq} - S^{**}_{,qq} \geq 0 \tag{2.5.30}$$

which would guarantee the inequality (2.5.16). This required bounds on q^{**} and ultimately, bounds on q. Thus the attempt failed due to circular reasoning. The conjecture that q is bounded holds only if q can be shown to be bounded by some independent means. It seems likely that (2.5.16) will have to be shown in an integral sense and not in a pointwise sense if this conjecture is to be upgraded to a proof.

In the remainder of this work the conjecture (2.5.6) is assumed to hold. This means that the satisfaction of a global entropy inequality implies nonlinear stability. The satisfaction of a cell-by-cell entropy inequality is certainly sufficient to satisfy a global inequality as well.

Chapter 3. THE SEMI-DISCRETE CASE

The previous chapter discussed conditions for L_2 bounds on the solution of systems of hyperbolic conservation laws. These conditions and the equations themselves assume an arbitrary distribution of q over the domain. Such a distribution requires an infinite amount of information to express, a feature that is not conducive to computational solution techniques. To avoid such a problem, this chapter will derive equations which are discrete in space but continuous in time. These are called the semi-discrete equations. Later, it will discuss a semi-discrete form of the second law and show that the conjecture of the previous chapter is not affected by the semi-discrete approximation. Finally, it will explore ways to construct numerical schemes which satisfy the semi-discrete equations while meeting semi-discrete constraints.

3.1 Deriving the Semi-Discrete Equations

As a first step, assume that the domain of interest is tessellated into a finite number of control volumes, each of which have a finite size. These can be identified by a single index j. Furthermore the surface bounding the j^{th} volume is denoted ∂V_j.

The integral equation describing such a volume is

$$\int_{V_j} q[t+\Delta t]\,dv - \int_{V_j} q[t]\,dv + \int_t^{t+\Delta t} \oint_{\partial V_j} \mathbf{f}\cdot\mathbf{n}\,dA\,d\tau = 0 \tag{3.1.1}$$

Because each control volume has a finite extent, no singularities can occur in the limit as $\Delta t \to 0$. Proceeding to this limit and dividing by Δt,

$$\frac{\mathrm{d}}{\mathrm{dt}} \int_{V_j} q\,dv + \oint_{\partial V_j} \mathbf{f}\cdot\mathbf{n}\,dA = 0 \tag{3.1.2}$$

The time indexes have been dropped because (3.1.2) holds at each time. At this point, the equations can be simplified by a definition.

$$q_j \equiv \frac{1}{V_j}\int_{V_j} q\,dv = \overline{q}_j \tag{3.1.3}$$

In one dimension there is a particularly simple numbering in which adjacent cells have adjacent numbers as shown in figure 3.1.

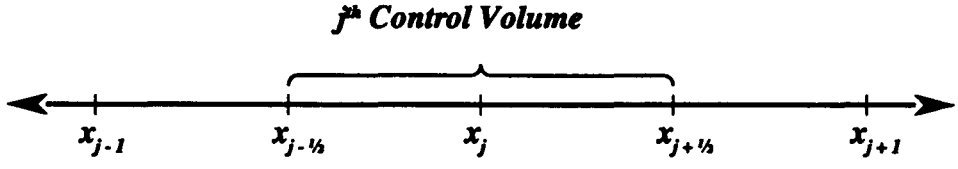

Figure 3.1. – A one-dimensional domain.

The surface integrals in one dimension are easily computed. Let the surface separating the j^{th} cell and the $j+1^{th}$ cell be denoted $j+\frac{1}{2}$, so that ∂V_j is composed of the two points $x_{j+\frac{1}{2}}$ and $x_{j-\frac{1}{2}}$. Finally, let $f \equiv \mathbf{f} \cdot \mathbf{i}$, where $\mathbf{i}$ is the unit vector in the x direction. This allows the surface integrals to be carried out in closed form. For example

$$\oint_{\partial V_j} \mathbf{f} \cdot \mathbf{n}\, dA = f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}} \tag{3.1.4}$$

In summary, the integral equations in one dimension reduce to

$$v_j(q_j)_{,t} + f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}} = 0 \tag{3.1.5}$$

Here the comma notation has been used to indicate differentiation. Traditionally the volume is divided out. In one dimension it seems a little silly to keep calling it a volume so $v_j = \Delta x_j$. This results in the semi-discrete form of (3.1.1).

$$(q_j)_{,t} + \frac{f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}}{\Delta x_j} = 0 \tag{3.1.6}$$

This equation is exact, as written. Using it requires computing $f_{j+\frac{1}{2}}$ and $f_{j-\frac{1}{2}}$. Assuming this is possible, the result of a time integration is a collection of values q_j at some future time. Usually what is needed is $q(x)$ instead, so some assumption needs to be made about the distribution of q within each volume.

Consider integration of (3.1.6) over a time sufficiently small that q_j is essentially constant. Then it is important to be sure that reconstruction of $q(x)$ from q_j is done in such a way that entropy is not destroyed. From this point of view the safest assumption is the piecewise constant distribution of Godunov.

$$q(x) = \overline{q}_j \qquad x_{j-\frac{1}{2}} \le x \le x_{j+\frac{1}{2}} \tag{3.1.7}$$

As shown by (2.2.6), this distribution provides the maximum entropy in each cell, consistent with the known values of q_j.

Starting with (2.1.7), a semi-discrete version of the second law can be constructed as well. This is

$$(\dot{P}_s)_j \equiv (S_j)_{,t} + \frac{F_{j+\frac{1}{2}} - F_{j-\frac{1}{2}}}{\Delta x_j} \ge 0 \tag{3.1.8}$$

The cell average entropy density, S_j, is defined

$$S_j \equiv \frac{1}{V_j} \int_{V_j} S(q)\, dv \tag{3.1.9}$$

Because of the assumed piecewise constant distribution of q,

$$S_j = S(\bar{q}_j) = S(q_j) \tag{3.1.10}$$

and the chain rule is applicable. In particular

$$(S_{,t})_j = (S_{,q})_j (q_{,t})_j \tag{3.1.11}$$

Substituting this into (3.1.8), using (3.1.6) to eliminate $(q_{,t})_j$, and multiplying through by Δx_j gives

$$\Delta x_j (\dot{P}_e)_j = -(S_{,q})_j (f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}) + (F_{j+\frac{1}{2}} - F_{j-\frac{1}{2}}) \geq 0 \tag{3.1.12}$$

This semi-discrete form of the second law has no time derivatives and can be computed from data at a single time level. The question of computing such quantities as $f_{j+\frac{1}{2}}$ and $F_{j+\frac{1}{2}}$ must now be addressed. Reviewing their origin (3.1.4), it follows that

$$f_{j+\frac{1}{2}} \equiv f(q_{j+\frac{1}{2}}) \tag{3.1.13}$$

$$F_{j+\frac{1}{2}} \equiv F(q_{j+\frac{1}{2}}) \tag{3.1.14}$$

where $q_{j+\frac{1}{2}}$ is the state which exists on the boundary between cells j and $j+1$. In virtually all of the work which follows, $q_{j+\frac{1}{2}}$ is approximated from q_j and q_{j+1}. Definitions (3.1.13) and (3.1.14) are then used to compute $f_{j+\frac{1}{2}}$ and $F_{j+\frac{1}{2}}$.

The definition of $f_{j+\frac{1}{2}}$ and $F_{j+\frac{1}{2}}$ from a common value of $q_{j+\frac{1}{2}}$, as in (3.1.13) and (3.1.14), differs from the usual practice. Usually $f_{j+\frac{1}{2}}$ is formed directly from $f(q_j)$ and $f(q_{j+1})$ without ever computing $q_{j+\frac{1}{2}}$. The quantity $F_{j+\frac{1}{2}}$ is approximated in some similar fashion. The author feels that the independent approximation of $F_{j+\frac{1}{2}}$ and $f_{j+\frac{1}{2}}$ does not do justice to the second law. Such an approach generates values of $F_{j+\frac{1}{2}}$ which can be grossly inaccurate in the neighborhood of shocks, where F is discontinuous. This can result in local anomalies such as oscillations. However, because of the global cancellation of entropy fluxes, stability is not affected by such errors. Support for this point of view is provided in Chapter 9, with a discussion of Tadmor's scheme.

It may seem that the piecewise constant assumption (3.1.7) precludes a unique value of $q_{j+\frac{1}{2}}$. This can be resolved by relaxing it slightly, in the immediate vicinity of $x_{j+\frac{1}{2}}$, to make $q(x)$ continuous. The modified assumption would be

$$q(x) = \bar{q}_j \qquad x_{j-\frac{1}{2}} + \epsilon \leq x \leq x_{j+\frac{1}{2}} - \epsilon \tag{3.1.15}$$

where $\epsilon << \Delta x_j$ is small enough that the integrals are not affected and the chain rule still holds.

Because $q(x)$ is continuous, any analysis valid for the integral equations is also valid for the semi-discrete equations. Due to the piecewise constant assumption, the domain integrals can easily be carried out cell-by-cell as sums. The conjecture is repeated here in the semi-discrete form.

CONJECTURE.

Given

$$f_{J+\frac{1}{2}} = f_{\frac{1}{2}} \tag{3.1.16}$$

$$\sum_{j=1}^{J} (q_j)_{,t} + f_{J+\frac{1}{2}} - f_{\frac{1}{2}} = 0 \tag{3.1.17}$$

$$F_{J+\frac{1}{2}} = F_{\frac{1}{2}} \tag{3.1.18}$$

$$\sum_{j=1}^{J} (S_j)_{,t} + F_{J+\frac{1}{2}} - F_{\frac{1}{2}} \geq 0 \tag{3.1.19}$$

$$\overline{q} \equiv \frac{\sum_{j=1}^{J} q_j \Delta x_j}{\sum_{j=1}^{J} \Delta x_j} \tag{3.1.20}$$

There exists a value q^* which can be determined from q such that

$$0 \leq \sum_{j=1}^{J} (q_j - \overline{q}_j)^T S^*_{,qq} (q_j - \overline{q}_j) \leq \sum_{j=1}^{J} S(\overline{q}) - S(q_j) \tag{3.1.21}$$

The terms having to do with initial conditions have been omitted here to avoid cluttering the notation. Otherwise there is no problem with including them. The constraint (3.1.19) can be satisfied with the cellwise condition

$$(\dot{P}_s)_j \geq 0 \tag{3.1.22}$$

where $(\dot{P}_s)_j$ was defined in (3.1.12).

The major change, in going from the integral form to the semi-discrete form of the equations, is the piecewise constant assumption (3.1.7). This means that $q(x)$ is constrained to be piecewise constant for all time. This does not affect the conservation laws or the second law of thermodynamics. Stability does not seem to be affected either, since the stability conjecture of the previous chapter continues to hold. There is, however, one major area where this constraint plays a role. That area is accuracy.

3.2 Accuracy in the Fine Mesh Limit

So far, the claim has been made that schemes which are conservative and satisfy the second law will be nonlinearly stable. In spite of this desirable property, such schemes are not useful unless they converge to the correct solution. One scheme which does not, is

$$q_j^{n+1} = q_j^n \tag{3.2.1}$$

This scheme conserves everything and ${q_j}^\infty = {q_j}^0$. Unfortunately, ${q_j}^\infty$ is generally incorrect.

At a minimum, schemes must be consistent in the sense of Lax if they are to converge to the correct solution. That is, if a numerical flux function, h, is defined such that

$$f_{j+\frac{1}{2}} = h(q_{j-1}, q_j, q_{j+1}, q_{j+2}) \tag{3.2.2}$$

then the consistency condition is

$$h(q, q, q, q) = f(q) \tag{3.2.3}$$

i.e., if the solution is a constant, the computed flux should have no error. This amounts to accuracy in the lowest order Taylor series sense.

Consider schemes of the form

$$f_{j+\frac{1}{2}} = f(q_{j+\frac{1}{2}}) \tag{3.2.4}$$

where

$$q_{j+\frac{1}{2}} = q_{j+\frac{1}{2}}(q_j, q_{j+1}) \tag{3.2.5}$$

Consistency for such schemes immediately follows if

$$q_{j+\frac{1}{2}}(q, q) = q \tag{3.2.6}$$

This condition will be used in most of the work that follows.

The question of accuracy remains. Traditionally, this question has been approached from the point of view of a Taylor series. To simplify the analysis, consider the scalar case. For example, (3.2.6) becomes

$$u_{j+\frac{1}{2}}(u, u) = u \tag{3.2.7}$$

This work uses interpolations of the form

$$u_{j+\frac{1}{2}} = u_j + \frac{1}{2}\phi_{j+\frac{1}{2}}(u_{j+1} - u_j) \tag{3.2.8}$$

where $\phi_{j+\frac{1}{2}}$ is an adjustable parameter, typically $0 \leq \phi_{j+\frac{1}{2}} \leq 2$. Interpolations such as (3.2.8) are used to form derivative approximations such as

$$\Delta x(u_{,x})_j \approx u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}} \tag{3.2.9}$$

The expression for $u_{j-\frac{1}{2}}$ is a simple index shift of (3.2.8). Substituting into (3.2.9) gives

$$\begin{aligned} \Delta x(u_{,x})_j \approx & \left[u_j \quad + \frac{1}{2}\phi_{j+\frac{1}{2}}(u_{j+1} - u_j)\right] \\ & - \left[u_{j-1} + \frac{1}{2}\phi_{j-\frac{1}{2}}(u_j - u_{j-1})\right] \end{aligned} \tag{3.2.10}$$

Assuming a constant Δx, and expanding the terms u_{j+1} and u_{j-1} about x_j gives the formula

$$\begin{aligned} \frac{u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}}{\Delta x} = & (u_{,x})_j \left[1 + \frac{1}{2}(\phi_{j+\frac{1}{2}} - \phi_{j-\frac{1}{2}})\right] \\ & - \frac{\Delta x}{2}(u_{,xx})_j \left[1 - \frac{1}{2}(\phi_{j+\frac{1}{2}} + \phi_{j-\frac{1}{2}})\right] \\ & + \frac{\Delta x^2}{6}(u_{,xxx})_j \left[1 + \frac{1}{2}(\phi_{j+\frac{1}{2}} - \phi_{j-\frac{1}{2}})\right] \\ & + O(\Delta x^3) \end{aligned} \tag{3.2.11}$$

At first glance, consistency appears to require that $\phi_{j+\frac{1}{2}} = \phi_{j-\frac{1}{2}}$. By induction, this implies a constant value of ϕ. By the same line of reasoning, second-order accuracy seems to require that the second term vanish, which happens if $\phi_{j+\frac{1}{2}} + \phi_{j-\frac{1}{2}} = 2$. Thus $\phi = 1$, the symmetric case, appears to be the only choice.

Fortunately, appearances can be deceiving. Less restrictive conditions for second-order accuracy can be written

$$\phi_{j+\frac{1}{2}} - \phi_{j-\frac{1}{2}} = O(\Delta x^2) \tag{3.2.12}$$

$$1 - \frac{1}{2}(\phi_{j+\frac{1}{2}} + \phi_{j-\frac{1}{2}}) = O(\Delta x) \tag{3.2.13}$$

This point of view allows a much broader class of second-order accurate schemes to be considered. For example, consider the choice

$$\phi_{j+\frac{1}{2}} = r_j \equiv \frac{u_j - u_{j-1}}{u_{j+1} - u_j} \tag{3.2.14}$$

It may be verified that

$$\phi_{j+\frac{1}{2}} = r_j \quad = 1 - \frac{(u_{,xx})_j}{(u_{,x})_j}\Delta x + O(\Delta x^2) \tag{3.2.15}$$

$$\phi_{j-\frac{1}{2}} = r_{j-1} = 1 - \frac{(u_{,xx})_j}{(u_{,x})_j}\Delta x + O(\Delta x^2) \tag{3.2.16}$$

although $r_j \neq r_{j-1}$. By inspection, this scheme meets conditions (3.2.12) and (3.2.13) and is therefore second-order accurate. This may be independently confirmed by substitution of (3.2.14) into (3.2.10). Dividing through by Δx yields the well-known approximation

$$\begin{aligned} \frac{u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}}{\Delta x} &= \frac{1}{\Delta x}\left[\frac{3}{2}u_j - 2u_{j-1} + \frac{1}{2}u_{j-2}\right] \\ &= (u_{,x})_j + O(\Delta x^2) \end{aligned} \tag{3.2.17}$$

Conditions (3.2.12) and (3.2.13) allow some freedom in the selection of $\phi_{j+\frac{1}{2}}$ and $\phi_{j-\frac{1}{2}}$. This freedom will be exploited in succeeding chapters to produce second-order accurate schemes which satisfy the second law.

3.3 Optimal Accuracy – Nonlinear Programming

Any time the term optimal is used, it must be carefully defined. Section 3.1 defines a cell entropy inequality (3.1.22) which is sufficient to guarantee nonlinear stability, at least for scalars. Equation (3.2.6) identifies a further constraint (consistency) which is required for convergence to the correct solution. Finally, equation (3.2.8) provides the freedom to achieve second-order accuracy in spite of these constraints. An optimally accurate scheme is defined here as the one which

(1) has the smallest possible entropy production rate and
(2) satisfies (3.1.22) and (3.2.6).

When q is a scalar, the selection of n independent values $(q_{j+\frac{1}{2}})$ is sufficient to determine a scheme. For optimal accuracy, these n unknowns must minimize a nonlinear functional while satisfying $2n-1$ coupled, nonlinear constraints. This is known as a nonlinear programming problem. Such problems are thought to be *NP* complete which is to say, they are completely intractable for any reasonable value of n.

Due to recent developments, the linear programming problem (linear functionals and constraints) is no longer *NP* complete. Solutions for n up to 1000 are now feasible. This is still not sufficient for our purpose in more than one dimension, even if the nonlinearity could be accommodated.

To make the problem more tractable, (while waiting for further advances in solution of the nonlinear programming problem) strict optimality can be given up in exchange for uncoupling the constraints. The following section demonstrates how this might be done.

3.4 First-Order Accurate Schemes

The first challenge in this nonlinear optimization problem is to show that the constraints can be met. This section will prove that they always can be, at least in the scalar case.

One constraint is the semi-discrete cell entropy inequality. This constraint is nonlinear and is applied at each control volume. Consider here the semi-discrete form of the second law in one dimension.

$$(S_j)_{,t} + \frac{F_{j+\frac{1}{2}} - F_{j-\frac{1}{2}}}{\Delta x_j} \geq 0 \tag{3.4.1}$$

This was originally derived in Section 3.1. There it was also shown (3.1.12) that an equivalent statement is

$$-(S_{,q})_j(f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}) + (F_{j+\frac{1}{2}} - F_{j-\frac{1}{2}}) \geq 0 \tag{3.4.2}$$

Equation (3.4.2) is algebraically identical to

$$\left[-(S_{,q})_j(f_{j+\frac{1}{2}} - f_j) + (F_{j+\frac{1}{2}} - F_j)\right] + \left[-(S_{,q})_j(f_j - f_{j-\frac{1}{2}}) + (F_j - F_{j-\frac{1}{2}})\right] \geq 0 \tag{3.4.3}$$

The left-hand side is an expression for the total rate of entropy production in the cell, $(\dot{P}_s)_j$. The terms in brackets define the right and left half cell entropy production rates, $(\dot{P}_s^+)_j$ and $(\dot{P}_s^-)_j$. The dependent variables here are $q_{j+\frac{1}{2}}$ (which defines $f_{j+\frac{1}{2}}$ and $F_{j+\frac{1}{2}}$) and $q_{j-\frac{1}{2}}$ (which defines $f_{j-\frac{1}{2}}$ and $F_{j+\frac{1}{2}}$). As written, $(\dot{P}_s^+)_j$ depends only on $q_{j+\frac{1}{2}}$, and $(\dot{P}_s^-)_j$ depends only on $q_{j-\frac{1}{2}}$. This suggests replacing (3.4.3) with the more restrictive constraints

$$(\dot{P}_s^+)_j \geq 0 \quad \text{and} \quad (\dot{P}_s^-)_j \geq 0 \tag{3.4.4}$$

Adoption of (3.4.4) allows $q_{j+\frac{1}{2}}$ to be selected independently of $q_{j-\frac{1}{2}}$. This decoupling of the constraints makes solution feasible.

Constraints (3.4.4) imply that $q_{j+\frac{1}{2}}$ must be chosen to simultaneously satisfy the following two inequalities

$$(\dot{P}_s^+)_j \geq 0 \quad \text{and} \quad (\dot{P}_s^-)_{j+1} \geq 0 \tag{3.4.5}$$

since (3.4.4) must be satisfied both at point j and point $j+1$. Is this always possible? For scalars, the answer is yes. For systems, the answer is not known in general. The argument follows.

Begin with an initial guess for $q_{j+\frac{1}{2}}$, a simple average of q_{j+1} and q_j. This allows $(\dot{P}_s^+)_j$ and $(\dot{P}_s^-)_{j+1}$ to be evaluated. If either of these is negative (the usual case), then $q_{j+\frac{1}{2}}$ will need to be adjusted. Say that $(\dot{P}_s^+)_j$ is negative. Then an appropriate change to $q_{j+\frac{1}{2}}$ is along the gradient, $\frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}}$. The differentiation is easily carried out

$$\frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}} = (F_{,q})_{j+\frac{1}{2}} - (S_{,q})_j(f_{,q})_{j+\frac{1}{2}} \tag{3.4.6}$$

Using the compatibility condition ($F_{,q} = S_{,q} f_{,q}$) this can be rewritten

$$\frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}} = \{(S_{,q})_{j+\frac{1}{2}} - (S_{,q})_j\}(f_{,q})_{j+\frac{1}{2}} \tag{3.4.7}$$

Adjusting $q_{j+\frac{1}{2}}$ in this direction will increase $(\dot{P}_s^+)_j$ as planned, but the same adjustment may decrease $(\dot{P}_s^-)_{j+1}$. This can be prevented if $q_{j+\frac{1}{2}}$ is restricted to values for which

$$\left(\frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}}\right) \cdot \left(\frac{\partial(\dot{P}_s^-)_{j+1}}{\partial q_{j+\frac{1}{2}}}\right) \geq 0 \tag{3.4.8}$$

It is instructive to carry out the differentiation and the dot product using (3.4.7).

$$\begin{aligned} &\left(\frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}}\right) \cdot \left(\frac{\partial(\dot{P}_s^-)_{j+1}}{\partial q_{j+\frac{1}{2}}}\right) \\ &\quad = \{(S_{,q})_{j+\frac{1}{2}} - (S_{,q})_j\}(f_{,q})_{j+\frac{1}{2}}(f_{,q})_{j+\frac{1}{2}}\{(S_{,q})_{j+1} - (S_{,q})_{j+\frac{1}{2}}\} \end{aligned} \tag{3.4.9}$$

Using Taylor's theorem with remainder

$$(S_{,q})_{j+\frac{1}{2}} - (S_{,q})_j = (q_{j+\frac{1}{2}} - q_j) S^*_{,qq} \tag{3.4.10}$$

(for some q^*) which allows (3.4.9) to be written as

$$\left(\frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}}\right) \cdot \left(\frac{\partial(\dot{P}_s^-)_{j+1}}{\partial q_{j+\frac{1}{2}}}\right) = (q_{j+\frac{1}{2}} - q_j) S^*_{,qq} (f_{,q})_{j+\frac{1}{2}} (f_{,q})_{j+\frac{1}{2}} S^{**}_{,qq} (q_{j+1} - q_{j+\frac{1}{2}}) \tag{3.4.11}$$

If q is a scalar quantity, then $S^*_{,qq}$ and $S^{**}_{,qq}$ are negative scalars (from convexity). Furthermore, $(f_{,q})_{j+\frac{1}{2}}$ is a scalar. It follows that condition (3.4.8) is satisfied for scalar q, if and only if

$$(q_{j+1} - q_{j+\frac{1}{2}})(q_{j+\frac{1}{2}} - q_j) \geq 0 \tag{3.4.12}$$

In words, this says that $q_{j+\frac{1}{2}}$ must lie between q_j and q_{j+1}. Within this range, any adjustment to $q_{j+\frac{1}{2}}$ which increases $(\dot{P}_s^+)_j$ will not decrease $(\dot{P}_s^-)_{j+1}$.

As long as $\frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}}$ does not change sign, $q_{j+\frac{1}{2}}$ can be adjusted all the way to q_j at which value $(\dot{P}_s^+)_j = 0$ by inspection. If $(\dot{P}_s^-)_{j+1}$ is negative, $q_{j+\frac{1}{2}}$ can be adjusted to q_{j+1}. In either case (3.4.1) will be satisfied.

Under what conditions can $\frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}}$ (and $\frac{\partial(\dot{P}_s^-)_{j+1}}{\partial q_{j+\frac{1}{2}}}$) change sign? Substituting 3.4.12 into 3.4.9 gives

$$\frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}} = (q_{j+\frac{1}{2}} - q_j) S^*_{,qq} (f_{,q})_{j+\frac{1}{2}} \tag{3.4.13}$$

Because of convexity, $S^*_{,qq}$ is a negative constant. Because of (3.4.12), $q_{j+\frac{1}{2}} - q_j$ has the same sign as $q_{j+1} - q_j$. Only $(f_{,q})_{j+\frac{1}{2}}$ can change sign on the interval. If this happens, there will be a value of $q_{j+\frac{1}{2}}$ on the interval, at which $(f_{,q})_{j+\frac{1}{2}} = 0$. If this represents a local maximum (sonic point), both $(\dot{P}^+_s)_j$ and $(\dot{P}^-_s)_{j+1}$ will be positive there. If it represents a local minimum (shock), $q_{j+\frac{1}{2}}$ will be adjusted away from it, to the appropriate endpoint value.

As a kind of side effect, the consistency constraints have been satisfied. Notice that (3.4.12) is sufficient to satisfy (3.2.6), which itself is sufficient for consistency in the sense of Lax.

A typical first-order upwind scheme might use the formula

$$q_{j+\frac{1}{2}} = q_j + \frac{1}{2}(1 - \text{sign}((f_{,q})_j + (f_{,q})_{j+1}))(q_{j+1} - q_j) \tag{3.4.14}$$

everywhere. The only change proposed here is the treatment of sonic points. At these points the formula

$$q_{j+\frac{1}{2}} = q_{\text{sonic}} \tag{3.4.15}$$

is used, where q_{sonic} is the value at which $(\dot{P}^+_s)_j$ (and $(\dot{P}^-_s)_{j+1}$) reaches a maximum.

The scheme described by (3.4.14) and (3.4.15) is first-order accurate in the limit of mesh refinement. The value ϕ referred to in Section 3.2 will always be 0 or 2 (except at sonic points), independent of the value of r. To achieve second-order accuracy, the scheme must be modified in such a way that $\phi \to 1$ as $r \to 1$. The next section will show how this can be done.

3.5 Second-Order Accuracy

In the previous section, an expression for the semi-discrete cell entropy inequality was derived. This is restated

$$(\dot{P}_s)_j \equiv (\dot{P}^-_s)_j + (\dot{P}^+_s)_j \geq 0 \tag{3.5.1}$$

The quantities $(\dot{P}^-_s)_j$ and $(\dot{P}^+_s)_j$ were defined shortly after (3.4.3). In Section 3.4 the constraint (3.5.1) was satisfied term by term. This is not necessary.

Begin as before by choosing, as an initial guess for $q_{j+\frac{1}{2}}$, a simple average of q_j and q_{j+1}. When this is done, trial values of $(\dot{P}^-_s)_j$ and $(\dot{P}^+_s)_j$ can be computed for each control volume. For illustration, suppose that $(\dot{P}^-_s)_j = 10$ and $(\dot{P}^+_s)_j = -9$. The first-order scheme in Section 3.4 would adjust $q_{j+\frac{1}{2}}$ until $q_{j+\frac{1}{2}} = q_j$, in order to assure that $(\dot{P}^+_s)_j \geq 0$. The resulting entropy production for the cell, $(\dot{P}_s)_j$, would then be at least 10. In fact, no adjustment is necessary. Using the trial values, $(\dot{P}_s)_j = 1$ which meets (3.5.1).

On the other hand, suppose that $(\dot{P}_s^+)_j = -11$ instead of -9. This would have no effect on the first-order scheme of the previous section. The second order scheme would find it sufficient to adjust $q_{j+\frac{1}{2}}$ very slightly, so that $(\dot{P}_s^+)_j = -10$ instead of the trial value of -11. With this change, $(\dot{P}_s)_j = 0$, which satisfies (3.5.1).

In both of these cases, satisfying (3.5.1) as a whole instead of term by term results in an order of magnitude reduction of the cell entropy production rate (dissipation). As the grid is refined, the trial values of $(\dot{P}_s^-)_j$ and $(\dot{P}_s^+)_j$ come closer and closer to satisfying (3.5.1). The needed adjustments get smaller and smaller. In the limit as the mesh is refined, the adjustments become vanishingly small, and $q_{j+\frac{1}{2}} \to \frac{1}{2}(q_j + q_{j+1})$. In this limit the scheme becomes second-order accurate.

Finally, suppose that $(\dot{P}_s^+)_j = -10$ and $(\dot{P}_s^-)_j = 0$. In this case, $q_{j+\frac{1}{2}}$ must be adjusted until $(\dot{P}_s^+)_j = 0$. In this case (which generally occurs at extrema) the large change in $q_{j+\frac{1}{2}}$ cannot be avoided. The scheme reverts to first-order accuracy at such points. Generally the number of extrema in the solution is small and does not increase with the number of mesh points.

A convincing demonstration of second-order accuracy requires a Taylor series expansion (without remainder this time) of $(\dot{P}_s)_j$ about the point j. This immediately leads (after dropping the constant terms) to

$$\begin{aligned}(\dot{P}_s)_j \approx &(F_{,q} - S_{,q}f_{,q})(q_{j+\frac{1}{2}} - q_j) + (F_{,q} - S_{,q}f_{,q})(q_{j-\frac{1}{2}} - q_j) \\ &+\frac{1}{2}(q_{j+\frac{1}{2}} - q_j)^T(F_{,qq} - S_{,q}f_{,qq})(q_{j+\frac{1}{2}} - q_j) \\ &-\frac{1}{2}(q_{j-\frac{1}{2}} - q_j)^T(F_{,qq} - S_{,q}f_{,qq})(q_{j-\frac{1}{2}} - q_j) \geq 0\end{aligned} \tag{3.5.2}$$

where everything is evaluated at point j unless otherwise noted. The first two terms vanish because of the compatability condition, $F_{,q} = S_{,q}f_{,q}$. Using this relation allows the second two terms to be simplified by expanding $F_{,qq}$. The simpler form is

$$(\dot{P}_s)_j \approx (q_{j+\frac{1}{2}} - q_j)^T(S_{,qq}f_{,q})(q_{j+\frac{1}{2}} - q_j) - (q_{j-\frac{1}{2}} - q_j)^T(S_{,qq}f_{,q})(q_{j-\frac{1}{2}} - q_j) \geq 0 \tag{3.5.3}$$

Using (3.2.8), the unknown values of $q_{j+\frac{1}{2}}$ can be replaced with unknown values of $\phi_{j+\frac{1}{2}}$ and $\phi_{j-\frac{1}{2}}$. After multiplying by 4 this gives

$$\begin{aligned}(\dot{P}_s)_j \approx &(q_{j+1} - q_j)^T(S_{,qq}f_{,q})\phi_{j+\frac{1}{2}}{}^2(q_{j+1} - q_j) - \\ &(q_{j-1} - q_j)^T(S_{,qq}f_{,q})(2 - \phi_{j-\frac{1}{2}})^2(q_{j-1} - q_j) \geq 0\end{aligned} \tag{3.5.4}$$

Using (3.2.11), the quantity $(q_{j-1} - q_j)$ can be replaced with $r_j(q_{j+1} - q_j)$. Collecting terms (valid for scalars) gives

$$(\dot{P}_s)_j \approx (q_{j+1} - q_j)^2 S_{,qq}f_{,q}(\phi_{j+\frac{1}{2}}{}^2 - r_j{}^2(2 - \phi_{j-\frac{1}{2}})^2) \geq 0 \tag{3.5.5}$$

The initial guesses for $q_{j+\frac{1}{2}}, q_{j-\frac{1}{2}}$ correspond to $\phi_{j+\frac{1}{2}} = \phi_{j-\frac{1}{2}} = 1$. As the mesh is refined, $r_j \rightarrow 1$. As this limit is approached, the values of ϕ will require less and less correction from the initial guess. Their final values will approach one. This is the condition given in Section 3.2 for second-order accuracy in the limit of mesh refinement.

In most solutions, there will be some regions where the second derivative cannot be neglected and r will not approach 1, no matter how much the mesh is refined. Typically these include discontinuities and, in some cases, extrema. At such places, a Taylor series shows first-order accuracy.

The second-order scheme differs from the first-order scheme in that (3.5.1) is satisfied as a whole instead of term by term. Shocks and sonic points are not affected. For the present, the scalar case can be considered complete. Chapters 5 and 6 give examples and results for some common scalar test problems. In Chapter 7, the formal extension to the Euler equations will be given.

Chapter 4. THE EFFECT OF TIME ADVANCE SCHEMES ON ENTROPY PRODUCTION RATES

In discussing the effects of a finite time step it may help to recall a similar problem for the incompressible Navier-Stokes equations. When these were first being solved, there was considerable difficulty with stability. Using spatial difference schemes with very high order spatial accuracy does not help. One solution to the stability dilemma is to construct schemes which satisfy a global inequality for kinetic energy. Except for the effect of boundary conditions, kinetic energy decreases steadily. Experience with incompressible Navier-Stokes solvers is instructive because kinetic energy is almost a formal entropy for these equations. Generally, such analysis is semi-discrete; i.e., it is done under the assumption that the time advance is analytic. This has seemed to work fairly well, even though time advance is generally not analytic.

Now the difference between a semi-discrete scheme (continuous in time) and certain fully discrete schemes (discrete in time) can be formally analyzed. In particular, the entropy production rates can be bounded on a cell-by-cell basis for both explicit Euler and implicit Euler time advance. For scalar equations, a modified form of Crank-Nicolson can also be analyzed.

4.1 The Fully Discrete Case

In the previous chapter, the integral equations were given on a volume-by-volume basis.

$$\int_{V_j} q[t+\Delta t]\,dv - \int_{V_j} q[t]\,dv + \int_t^{t+\Delta t} \oint_{\partial V_j} \mathbf{f}\cdot\mathbf{n}\,dA\,d\tau = 0 \tag{4.1.1}$$

To simplify notation, time is indexed by a superscript so that

$$q_j^n \equiv q_j[t] \qquad q_j^{n+1} \equiv q_j[t+\Delta t] \tag{4.1.2}$$

Using the definitions (3.1.3) and (3.1.4), (4.1.1) becomes

$$q_j^{n+1} - q_j^n + \int_t^{t+\Delta t} \frac{(f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}})}{\Delta x_j}\,d\tau = 0 \tag{4.1.3}$$

which is the fully discrete analog of (3.1.6). Like (3.1.6) it is exact. All that remains is to integrate the fluxes over the time interval. Since the entropy flux will also have to be integrated, a simple strategy is to assume that $q_{j+\frac{1}{2}}$ remains constant over the entire time interval. For example, explicit Euler time advance is equivalent to

$$q_{j+\frac{1}{2}}[\tau] = q_{j+\frac{1}{2}}^n \quad \text{for all} \quad n\Delta t \le \tau < (n+1)\Delta t \tag{4.1.4}$$

and implicit Euler is equivalent to

$$q_{j+\frac{1}{2}}[\tau] = q_{j+\frac{1}{2}}^{n+1} \quad \text{for all} \quad n\Delta t < \tau \leq (n+1)\Delta t \tag{4.1.5}$$

With the addition of definitions like

$$f_{j+\frac{1}{2}}^{n} \equiv f(q_{j+\frac{1}{2}}^{n}) \tag{4.1.6}$$

the fully discrete form of the equations becomes, for explicit Euler

$$q_j^{n+1} - q_j^n + \frac{\Delta t}{\Delta x_j}(f_{j+\frac{1}{2}}^n - f_{j-\frac{1}{2}}^n) = 0 \tag{4.1.7}$$

Using a similar derivation, the fully discrete form of the second law becomes

$$S_j^{n+1} - S_j^n + \frac{\Delta t}{\Delta x_j}(F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n) \geq 0 \tag{4.1.8}$$

Equations (4.1.7) and (4.1.8) will be used in the next section to show that schemes using explicit Euler time advance have a lower (perhaps negative) entropy production rate than the same scheme with analytic time advance. This is true, cell by cell, for any positive value of Δt.

4.2 Explicit Euler Time Advance

This section explores the irreversibility of explicit Euler time advance. The entropy production due to time advance can be separated from the entropy production due to space differencing. This is shown in the following theorem.

THEOREM.

Given a fully discrete scheme using explicit Euler time advance

$$(q_j^{n+1} - q_j^n) + \Delta t \frac{f_{j+\frac{1}{2}}^n - f_{j-\frac{1}{2}}^n}{\Delta x_j} = 0 \tag{4.2.1}$$

Then

$$(\dot{P}_s)_j^{EE} \leq (\dot{P}_s)_j \tag{4.2.2}$$

where

$$(\dot{P}_s)_j^{EE} \equiv \frac{S_j^{n+1} - S_j^n}{\Delta t} + \frac{F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n}{\Delta x_j} \tag{4.2.3}$$

and

$$(\dot{P}_s)_j \equiv (S_{,t})_j^n + \frac{F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n}{\Delta x_j}$$
$$= -(S_{,q})_j^n \frac{f_{j+\frac{1}{2}}^n - f_{j-\frac{1}{2}}^n}{\Delta x_j} + \frac{F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n}{\Delta x_j} \tag{4.2.4}$$

PROOF:

Even though (4.2.4) is semi-discrete, $q(x)$ must be specified. Here, $q(x)$ at time level n has been selected. The quantities S_j^{n+1} and S_j^n are related through Taylor's theorem with remainder:

$$S_j^{n+1} = S_j^n + (S_{,q})_j^n (q_j^{n+1} - q_j^n) + \frac{1}{2}(q_j^{n+1} - q_j^n)^T S_{,qq}^{\xi} (q_j^{n+1} - q_j^n) \tag{4.2.5}$$

where $n \leq \xi \leq n+1$. Equation (4.2.5) is exact, however the value of ξ is not known. The right-hand side of (4.2.5) may then be substituted for S_j^{n+1} in (4.2.3) to give

$$(\dot{P}_s)_j^{EE} = \frac{1}{\Delta t}\left((S_j^n + (S_{,q})_j^n (q_j^{n+1} - q_j^n) + \frac{1}{2}(q_j^{n+1} - q_j^n)^T S_{,qq}^{\xi} (q_j^{n+1} - q_j^n)) - S_j^n \right)$$
$$+ \frac{F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n}{\Delta x_j} \tag{4.2.6}$$

Furthermore, equation (4.2.1) can be used for the value of $(q_j^{n+1} - q_j^n)$. This gives

$$(\dot{P}_s)_j^{EE} = \left\{ -(S_{,q})_j^n \frac{f_{j+\frac{1}{2}}^n - f_{j-\frac{1}{2}}^n}{\Delta x_j} + \frac{F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n}{\Delta x_j} \right\}$$
$$+ \frac{1}{2\Delta t}(q_j^{n+1} - q_j^n)^T S_{,qq}^{\xi} (q_j^{n+1} - q_j^n) \tag{4.2.7}$$

The quantity in braces can be replaced with the semi-discrete entropy production rate through (4.2.4) to yield

$$(\dot{P}_s)_j^{EE} = (\dot{P}_s)_j + \frac{1}{2\Delta t}(q_j^{n+1} - q_j^n)^T S_{,qq}^{\xi} (q_j^{n+1} - q_j^n) \tag{4.2.8}$$

Since $S_{,qq}$ is a negative definite matrix (by convexity), the second term of (4.2.8) is a proper quadratic form and cannot be positive. Thus, the fully discrete entropy production rate $(\dot{P}_s)_j^{EE}$ must be equal to or less than the semi-discrete entropy production $(\dot{P}_s)_j$.

What does this say about explicit Euler as a time advance scheme? For problems which naturally create entropy (such as the heat equation), it may be a reasonable scheme.

Since $(q_j^{n+1} - q_j^n)$ is roughly proportional to Δt, the quadratic term is first-order in time. This agrees with existing analysis that explicit Euler is first-order accurate.

Stability depends on the relative size of the (presumably positive) semi-discrete entropy production rate, $(\dot{P}_s)_j$, and the (generally negative) quadratic term. For small enough time steps, $(\dot{P}_s)_j$ dominates, and the fully discrete entropy production rate is positive ($(\dot{P}_s)_j^{EE} \geq 0$). For very large time steps the opposite occurs. Thus the entropy approach demonstrates that explicit Euler is conditionally stable for dissipative systems. As a steady state is approached, $(q_j^{n+1} - q_j^n)$ vanishes and the entropy production rate is entirely determined by the space differencing, as expected.

Convection problems are the main concern in this work. Provided that the derivatives exist, such problems can be expressed as

$$q_{,t} + f_{,x} = 0 \tag{4.2.9}$$

If $q(x,t)$ is smooth enough to allow use of the chain rule, it can easily be shown that $\dot{P}_s = 0$, analytically. This was shown in Section 2.2.

The more accurate a space differencing scheme becomes, the closer it comes to this analytic limit. An ideal space differencing scheme, one with no error, has no semi-discrete entropy production, except at shocks. For such a scheme, explicit Euler is not stable for any positive time step Δt. The usual fix is to add artificial spatial dissipation, which modifies the equations. The resulting solutions can be surprisingly good for transient problems, if the spatial dissipation (an error) approximately cancels the temporal dissipation (an error of opposite sign). The first-order upwind scheme, for example, produces no error for the wave equation if $\Delta t = c\Delta x$ (where c is the local wave speed). This strategy of canceling errors is difficult in practice (for all but the simplest problems) because c is not constant in space or time and because there may be more than one characteristic speed. Furthermore, it doesn't work at steady state or at very small time steps because the time advance errors vanish, while the spatial errors which they are supposed to cancel, remain.

4.3 Implicit Euler Time Advance

This section explores the irreversibility of implicit Euler time advance. The entropy production due to time advance can be separated from the entropy production due to space differencing, just as for explicit Euler. This is shown in the following theorem:

THEOREM.

Given some fully discrete scheme using implicit Euler time advance

$$(q_j^{n+1} - q_j^n) + \Delta t \frac{f_{j+\frac{1}{2}}^{n+1} - f_{j-\frac{1}{2}}^{n+1}}{\Delta x_j} = 0 \tag{4.3.1}$$

Then

$$(\dot{P}_s)_j^{IE} \geq (\dot{P}_s)_j \tag{4.3.2}$$

where

$$(\dot{P}_s)_j^{IE} \equiv \frac{S_j^{n+1} - S_j^n}{\Delta t} + \frac{F_{j+\frac{1}{2}}^{n+1} - F_{j-\frac{1}{2}}^{n+1}}{\Delta x_j} \tag{4.3.3}$$

and

$$\begin{aligned}(\dot{P}_s)_j &\equiv (S_{,t})_j^{n+1} + \frac{F_{j+\frac{1}{2}}^{n+1} - F_{j-\frac{1}{2}}^{n+1}}{\Delta x_j} \\ &= -(S_{,q})_j^{n+1} \frac{f_{j+\frac{1}{2}}^{n+1} - f_{j-\frac{1}{2}}^{n+1}}{\Delta x_j} + \frac{F_{j+\frac{1}{2}}^{n+1} - F_{j-\frac{1}{2}}^{n+1}}{\Delta x_j}\end{aligned} \tag{4.3.4}$$

PROOF:

Even though (4.3.4) is semi-discrete, $q(x)$ must be specified. Here, $q(x)$ at time level $n+1$ has been selected. The quantities S_j^{n+1} and S_j^n are related through Taylor's theorem with remainder:

$$S_j^n = S_j^{n+1} - (S_{,q})_j^{n+1}(q_j^{n+1} - q_j^n) + \frac{1}{2}(q_j^{n+1} - q_j^n)^T S_{,qq}^{\eta}(q_j^{n+1} - q_j^n) \tag{4.3.5}$$

where $n \leq \eta \leq n+1$. Equation (4.3.5) is exact, however the value of η is not known. The right-hand side of (4.3.5) may now be substituted for S_j^{n+1} in (4.3.3) to give

$$\begin{aligned}(\dot{P}_s)_j^{IE} = \frac{1}{\Delta t}\Bigg(& S_j^{n+1} - \Big(S_j^{n+1} - (S_{,q})_j^{n+1}(q_j^{n+1} - q_j^n) \\ & + \frac{1}{2}(q_j^{n+1} - q_j^n)^T S_{,qq}^{\eta}(q_j^{n+1} - q_j^n)\Big)\Bigg) \\ & + \frac{F_{j+\frac{1}{2}}^{n+1} - F_{j-\frac{1}{2}}^{n+1}}{\Delta x_j}\end{aligned} \tag{4.3.6}$$

Furthermore, equation (4.3.1) can be used for the value of $(q_j^{n+1} - q_j^n)$. This gives

$$\begin{aligned}(\dot{P}_s)_j^{IE} = & \left\{ -(S_{,q})_j^{n+1} \frac{f_{j+\frac{1}{2}}^{n+1} - f_{j-\frac{1}{2}}^{n+1}}{\Delta x_j} + \frac{F_{j+\frac{1}{2}}^{n+1} - F_{j-\frac{1}{2}}^{n+1}}{\Delta x_j} \right\} \\ & + \frac{1}{2\Delta t}(q_j^{n+1} - q_j^n)^T S_{,qq}^{\eta}(q_j^{n+1} - q_j^n)\end{aligned} \tag{4.3.7}$$

The quantity in braces can be replaced with the semi-discrete entropy production rate through (4.3.4) to yield

$$(\dot{P}_s)_j^{IE} = (\dot{P}_s)_j - \frac{1}{2\Delta t}(q_j^{n+1} - q_j^n)^T S_{,qq}^{\eta}(q_j^{n+1} - q_j^n) \tag{4.3.8}$$

Since $S_{,qq}$ is a negative definite matrix (by convexity), the second term of 4.3.9 is a proper quadratic form. Thus, the fully discrete entropy production rate $(\dot{P}_s)_j^{IE}$ must be equal to or greater than the semi-discrete entropy production $(\dot{P}_s)_j$.

What does this say about implicit Euler as a time advance scheme? Since $(q_j^{n+1} - q_j^n)$ is roughly proportional to Δt, the quadratic term is first-order in time. This agrees with existing analysis that implicit Euler is first-order accurate. In this regard implicit Euler is very similar to explicit Euler.

In the area of stability, though, the two schemes are very different. Stability depends only on constructing a space differencing scheme for which $(\dot{P}_s)_j \geq 0$. It then follows that $(\dot{P}_s)_j^{IE} \geq 0$ also, independent of the time step. Thus the entropy approach demonstrates that implicit Euler is unconditionally stable *in a nonlinear sense* for convective or dissipative systems. The catch is that this wonderful property only follows if the scheme is carried out exactly. For most problems, this implies solving coupled systems of nonlinear equations, an expensive process. Solution of these systems is possible if the time step is small enough to allow linearization in q, however. Thus, while implicit Euler offers unconditional stability, this advantage is fully realized only in linear problems. For nonlinear problems, some time step restriction may result from linearization errors. This restriction relaxes as a steady solution is approached. Furthermore, at steady state, $(q_j^{n+1} - q_j^n)$ vanishes and the entropy production rate is entirely determined by the space differencing, as expected.

4.4 A Second-Order Accurate Time Advance Scheme

It is interesting to note the similarity in the analysis of explicit Euler and implicit Euler time advance. Implicit Euler is essentially the same as explicit Euler with a negative time step. The errors incurred (comparing (4.3.8) with (4.2.8)) are of the same form, but of opposite sign. This suggests that it might be possible, by combining the two methods, to construct a time advance having no temporal dissipation. In particular, if a half step of implicit Euler is followed by a step of explicit Euler, the temporal dissipation might be partially or completely eliminated.

THEOREM.

Given some fully discrete scheme; a time advance of the form

$$(q_j^{n+1} - q_j^n) + \Delta t \frac{f_{j+\frac{1}{2}}^{n+\frac{1}{2}} - f_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x_j} = 0 \tag{4.4.1}$$

which is composed of a half step of implicit Euler ($t \leq \tau \leq t + \frac{\Delta t}{2}$) followed by a half step of explicit Euler ($t + \frac{\Delta t}{2} \leq \tau \leq t + \Delta t$),

Then

$$(\dot{P}_s)_j^{MCN} = (\dot{P}_s)_j + \frac{1}{8\Delta t}(q_j^{n+1} - q_j^n)^T(S_{,qq}^{\xi} - S_{,qq}^{\eta})(q_j^{n+1} - q_j^n) \tag{4.4.2}$$

where the precise evaluation states of $S_{,qq}^{\xi}$ and $S_{,qq}^{\eta}$ are not known. Equation (4.4.2) uses the definitions

$$(\dot{P}_s)_j^{MCN} \equiv \frac{S_j^{n+1} - S_j^n}{\Delta t} + \frac{F_{j+\frac{1}{2}}^{n+\frac{1}{2}} - F_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x_j} \tag{4.4.3}$$

and

$$\begin{aligned}(\dot{P}_s)_j &\equiv (S_{,t})_j^{n+\frac{1}{2}} + \frac{F_{j+\frac{1}{2}}^{n+\frac{1}{2}} - F_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x_j} \\ &= -(S_{,q})_j^{n+\frac{1}{2}} \frac{f_{j+\frac{1}{2}}^{n+\frac{1}{2}} - f_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x_j} + \frac{F_{j+\frac{1}{2}}^{n+\frac{1}{2}} - F_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x_j}\end{aligned} \tag{4.4.4}$$

PROOF:

Here, $q(x)$ at time level $n+\frac{1}{2}$ has been selected for the semi-discrete terms. The quantities S_j^{n+1} and S_j^n are related, through Taylor's theorem with remainder to $S_j^{n+\frac{1}{2}}$:

$$S_j^n = S_j^{n+\frac{1}{2}} - (S_{,q})_j^{n+\frac{1}{2}}(q_j^{n+\frac{1}{2}} - q_j^n) + \frac{1}{2}(q_j^{n+\frac{1}{2}} - q_j^n)^T S_{,qq}^{\eta}(q_j^{n+\frac{1}{2}} - q_j^n) \tag{4.4.5}$$

$$S_j^{n+1} = S_j^{n+\frac{1}{2}} + (S_{,q})_j^{n+\frac{1}{2}}(q_j^{n+1} - q_j^{n+\frac{1}{2}}) + \frac{1}{2}(q_j^{n+1} - q_j^{n+\frac{1}{2}})^T S_{,qq}^{\xi}(q_j^{n+1} - q_j^{n+\frac{1}{2}}) \tag{4.4.6}$$

where $n \le \eta \le n+\frac{1}{2} \le \xi \le n+1$. Substituting (4.4.5) and (4.4.6) into (4.4.3) gives

$$\begin{aligned}(\dot{P}_s)_j^{MCN} = \frac{1}{\Delta t}\Big(&\big(S_j^{n+\frac{1}{2}} + (S_{,q})_j^{n+\frac{1}{2}}(q_j^{n+1} - q_j^{n+\frac{1}{2}}) \\ &\quad + \frac{1}{2}(q_j^{n+1} - q_j^{n+\frac{1}{2}})^T S_{,qq}^{\xi}(q_j^{n+1} - q_j^{n+\frac{1}{2}})\big) \\ &- \big(S_j^{n+\frac{1}{2}} - (S_{,q})_j^{n+\frac{1}{2}}(q_j^{n+\frac{1}{2}} - q_j^n) \\ &\quad + \frac{1}{2}(q_j^{n+\frac{1}{2}} - q_j^n)^T S_{,qq}^{\eta}(q_j^{n+\frac{1}{2}} - q_j^n)\big)\Big) \\ &+ \frac{F_{j+\frac{1}{2}}^{n+\frac{1}{2}} - F_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x_j}\end{aligned} \tag{4.4.7}$$

Collecting terms gives

$$(\dot{P}_s)_j^{MCN} = \frac{1}{\Delta t}\Bigg((S_{,q})_j^{n+\frac{1}{2}}(q_j^{n+1} - q_j^n) + \frac{1}{2}(q_j^{n+1} - q_j^{n+\frac{1}{2}})^T S_{,qq}^{\xi}(q_j^{n+1} - q_j^{n+\frac{1}{2}}) - \frac{1}{2}(q_j^{n+\frac{1}{2}} - q_j^n)^T S_{,qq}^{\eta}(q_j^{n+\frac{1}{2}} - q_j^n)\Bigg) + \frac{F_{j+\frac{1}{2}}^{n+\frac{1}{2}} - F_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x_j} \tag{4.4.8}$$

The key observation at this point is that

$$(q_j^{n+1} - q_j^{n+\frac{1}{2}}) = (q_j^{n+\frac{1}{2}} - q_j^n) = -\frac{\Delta t}{2}\frac{f_{j+\frac{1}{2}}^{n+\frac{1}{2}} - f_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x_j} = \frac{1}{2}(q_j^{n+1} - q_j^n) \tag{4.4.9}$$

Furthermore, equation (4.4.1) can be used for the value of $(q_j^{n+1} - q_j^n)$. This gives

$$(\dot{P}_s)_j^{MCN} = \left\{ -(S_{,q})_j^{n+\frac{1}{2}}\frac{f_{j+\frac{1}{2}}^{n+\frac{1}{2}} - f_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x_j} + \frac{F_{j+\frac{1}{2}}^{n+\frac{1}{2}} - F_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x_j} \right\} + \frac{1}{8\Delta t}(q_j^{n+1} - q_j^n)^T(S_{,qq}^{\xi} - S_{,qq}^{\eta})(q_j^{n+1} - q_j^n) \tag{4.4.10}$$

The quantity in braces can be replaced with the semi-discrete entropy production rate through (4.4.4) to yield

$$(\dot{P}_s)_j^{MCN} = (\dot{P}_s)_j + \frac{1}{8\Delta t}(q_j^{n+1} - q_j^n)^T(S_{,qq}^{\xi} - S_{,qq}^{\eta})(q_j^{n+1} - q_j^n) \tag{4.4.11}$$

which is the claimed result.

Although $S_{,qq}^{\xi}$ and $S_{,qq}^{\eta}$ are both negative definite, the sign of their difference is indeterminate. Although they are evaluated at states which may be different, both are, in a Taylor series sense, evaluated close to $q_j^{n+\frac{1}{2}}$. Their difference is of order Δt. This accounts for the second-order accuracy in time that this scheme enjoys under conventional analysis techniques.

The behavior of this time advance scheme for scalars is quite remarkable. As previously noted, $S_{,qq}^{\xi} = S_{,qq}^{\eta} = -2$ in this case. Equation (4.4.11) indicates that $(\dot{P}_s)_j^{MCN} = (\dot{P}_s)_j$; the scheme provides no temporal dissipation at all. This is true cell-by-cell. Thus, the discrete time advance is as good as an analytic time advance. This follows regardless of the time step.

Since half the scheme is implicit Euler, it is reasonable to expect that the linearization problems of implicit Euler will carry over. Indeed they do. The scheme is, however, no

more difficult to implement than implicit Euler. Given that one can solve for $q_j^{n+\frac{1}{2}}$ using implicit Euler, the equivalent higher order scheme is

$$q_j^{n+1} = q_j^n + 2(q_j^{n+\frac{1}{2}} - q_j^n) \tag{4.4.12}$$

This follows directly from (4.4.9). Many codes in use today use the so-called "delta" form. Equation (4.4.12) shows that, simply by doubling the correction (a sort of overrelaxation), it is possible to achieve second-order accuracy in time. This fact is apparently known, but seldom used.

There is a connection between this scheme and the more common Crank-Nicolson, time advance. Crank-Nicolson consists of a half step of explicit Euler, followed by a half step of implicit Euler. Though the scheme shown here reverses the two, a sequence of Crank-Nicolson steps preceded by a half step of implicit Euler and succeeded by a half step of explicit Euler will yield exactly the same result as the scheme suggested here.

Traditionally one includes, for generality, the so-called θ schemes in an analysis of this sort. These are variants where one advances in time by $\theta\Delta t$, using implicit Euler and then advances $(1-\theta)\Delta t$, using explicit Euler. Values of $\theta = 0, \frac{1}{2}$, and 1 correspond to explicit Euler, modified Crank–Nicolson, and implicit Euler. These are the only values ever used in practice. Suffice to say that the analysis shown in this section is easily extendable to other values of θ.

4.5 Practical Considerations for Implicit Schemes

In linear problems, it is possible to implement implicit schemes with style and grace, retaining rigor and exactitude. On the other hand, most problems of interest are nonlinear. What can be done in these cases? Usually all that can be done is to locally linearize and take small time steps. This section will address the question of how to maintain the entropy inequality while doing this.

Implementing implicit Euler time advance (eq. 4.3.1) requires evaluation of terms such as $f_{j+\frac{1}{2}}^{n+1}$. In general, this is very difficult. What is easy to evaluate is $f_{j+\frac{1}{2}}^n$. If the two fluxes are not too different, it is reasonable to use a Taylor series in place of the unknown quantity. Supposing, for illustration, that $f_{j+\frac{1}{2}}^n = f(q_j^n, q_{j+1}^n)$, then a natural approximation is

$$f_{j+\frac{1}{2}}^{n+1} \approx f_{j+\frac{1}{2}}^n + \frac{\partial f_{j+\frac{1}{2}}}{\partial q_j}(q_j^{n+1} - q_j^n) + \frac{\partial f_{j+\frac{1}{2}}}{\partial q_{j+1}}(q_{j+1}^{n+1} - q_{j+1}^n) \tag{4.5.1}$$

This approximation makes implicit methods possible. It reduces a coupled set of nonlinear algebraic equations to a coupled set of linear algebraic equations. Such systems are easy to

solve, expecially in one dimension. Typically a tridiagonal or pentadiagonal matrix needs to be solved.

How appropriate is this approximation? If $(q_j^{n+1} - q_j^n)$ and $(q_{j+1}^{n+1} - q_{j+1}^n)$ are "small enough," then it is an excellent approximation. Their size is, of course, directly related to the time step Δt. By controlling Δt we can make them "small enough." How small is that? Small enough that the discrete entropy production rate (e.g., eq. 4.3.8) remains positive.

Another problem appears. Even for a very small (finite) time step, there are linearization errors in the approximation (4.5.1). This introduces errors, even in an *a posteriori* calculation of the entropy production rate. These uncertainties jeopardize the conclusions of Section 4.3 (and Section 2.5).

The most straightforward solution is to iterate within each time step to exactly solve the nonlinear algebraic equations. This is expensive however and may not be robust due to certain roundoff difficulties.

A cheaper plan is an implicit-explicit scheme of the following type. First perform an approximate implicit step to get an approximation to q^{n+1}.

$$\mathbf{A}(\hat{q} - q^n) = \delta_x f \tag{4.5.2}$$

where $\mathbf{A}$ is the matrix implicit in (4.5.1) and $\delta_x f$ is a vector of flux differences. The vector $\hat{q}$ is approximately (exactly for linear problems) equal to q^{n+1}. Equation (4.5.2) allows solution for $\hat{q}$ at all grid points by standard linear algebra techniques.

Second, compute the fluxes for this approximate solution.

$$\hat{f}_{j+\frac{1}{2}} = f(\hat{q}_{j+\frac{1}{2}}) \tag{4.5.3}$$

The quantities $\hat{q}_{j+\frac{1}{2}}$ are computed from $\hat{q}_j$ and $\hat{q}_{j+1}$ using whatever semi-discrete scheme is chosen.

Finally, use these fluxes explicitly to compute q^{n+1}.

$$q_j^{n+1} = q_j^n + \frac{\Delta t}{\Delta x}(\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}) \tag{4.5.4}$$

This allows an *a posteriori* check on the entropy inequality everywhere.

$$S(q_j^{n+1}) - S(q_j^n) + \frac{\Delta t}{\Delta x}\left(F(\hat{q}_{j+\frac{1}{2}}) - F(\hat{q}_{j-\frac{1}{2}})\right) \geq 0 \tag{4.5.5}$$

If it fails anywhere, the time step is too large. With this test, one can be sure that approximation (4.5.1) is "good enough," at least in the sense that stability will be preserved. A similar strategy can be performed for the second-order accurate scheme of Section 4.4 through the implementation shown in equation (4.4.12).

Chapter 5. LINEAR SCALAR EQUATIONS

5.1 A Suitable Entropy for the Wave Equation

This chapter will consider solution of the ordinary scalar wave equation in one dimension. Despite its simplicity, this equation provides a good illustration of the principles involved. It is a reasonable model for a contact discontinuity, because in both cases the characteristics are parallel.

For this and other scalar equations the variable u will be used instead of q, mostly because these equations are usually written that way. The scalar wave equation in one dimension is

$$u_{,t} + cu_{,x} = 0 \tag{5.1.1}$$

where the constant c is the wave speed. The exact solution is a translation of the initial conditions with speed c. The boundary conditions are considered to be periodic. The exact solution has no dissipation; it conserves entropy. Numerical schemes satisfying the second law will usually produce some (hopefully small) amount of entropy. This results in smearing of the initial conditions.

The results of Chapter 4 leave only the task of finding a spatial differencing scheme which satisfies the semi-discrete entropy inequality. Using such a scheme with Modified Crank–Nicolson or Implicit Euler time advance produces a scheme satisfying the fully discrete entropy inequality.

Consider first the semi-discrete approximation.

$$u_{,t} + \frac{f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}}{\Delta x_j} = 0 \tag{5.1.2}$$

For this equation $f \equiv cu$. Since c is constant (5.1.2) reduces to

$$u_{,t} + \frac{c}{\Delta x}(u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}) = 0 \tag{5.1.3}$$

The idea here is to pick $u_{j+\frac{1}{2}}$ in such a way as to satisfy a cell entropy inequality. Before that can be done an entropy has to be defined. In Section 2.5 it was shown that $S \equiv -u^2$ is a suitable entropy for scalar equations such as this. The convexity condition is easily verified.

$$S_{,uu} = -2 \leq 0 \tag{5.1.4}$$

A corresponding entropy flux can be obtained by direct integration of the compatibility condition

$$F_{,u} = S_{,u} f_{,u} \tag{5.1.5}$$

The right-hand side evaluates to $-2cu$ which is easily integrated to find the entropy flux F. A suitable entropy pair then, is

$$S = -u^2 \qquad \text{and} \qquad F = -cu^2 \tag{5.1.6}$$

This choice allows a semi-discrete entropy inequality to be formulated. Substituting $S_{,u}$ for $S_{,q}$, cu for f and $-cu^2$ for F, the semi-discrete entropy inequality (3.1.12) is, after multiplication by Δx,

$$\Delta x(\dot{P}_s)_j = 2cu_j(u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}) - c\Big((u_{j+\frac{1}{2}})^2 - (u_{j-\frac{1}{2}})^2\Big) \geq 0 \tag{5.1.7}$$

Remarkably, this expression factors neatly into a difference of two squares.

$$\Delta x(\dot{P}_s)_j = -c\left[(u_{j+\frac{1}{2}} - u_j)^2 - (u_j - u_{j-\frac{1}{2}})^2\right] \geq 0 \tag{5.1.8}$$

This expression for the semi-discrete entropy production rate is exact. All that remains is to choose $u_{j+\frac{1}{2}}$ and $u_{j-\frac{1}{2}}$ in such a way that (5.1.8) is satisfied.

5.2 A First-Order Scheme

Construction of first-order accurate schemes which satisfy a cell entropy inequality was discussed in Section 3.4. The general idea is to eliminate negative terms from the entropy production rate. This is easily accomplished in this case. For positive values of c, the choice

$$u_{j+\frac{1}{2}} \equiv u_j \tag{5.2.1}$$

will leave only the positive term, $c(u_j - u_{j-\frac{1}{2}})^2$, in (5.1.8). If c happens to be negative, the definition $u_{j-\frac{1}{2}} \equiv u_j$ should be used instead, leaving the positive term $-c(u_{j+\frac{1}{2}} - u_j)^2$. Since c is a constant, never changing sign, there can be no conflicts requiring $u_{j+\frac{1}{2}}$ to equal both u_j and u_{j+1}.

The total entropy production rate for this scheme can easily be calculated. The entropy production for the j^{th} cell is

$$\Delta x(\dot{P}_s)_j = c(u_j - u_{j-\frac{1}{2}})^2 \tag{5.2.2}$$

for $c > 0$. By definition (5.2.1) $u_{j-\frac{1}{2}} = u_{j-1}$ in this case, so

$$\Delta x(\dot{P}_s)_j = c(u_j - u_{j-1})^2 \tag{5.2.3}$$

Thus the total entropy production is the sum of the squares of the differences between grid points, a result which holds even on irregular meshes. Since total entropy production should be zero, a reasonable measure of error can be obtained by summing $\Delta x(\dot{P}_s)_j$ over

the mesh. This measure of the total error will always be a positive quantity for nontrivial initial conditions. Assume for now that an "optimum" mesh has somehow been chosen, one in which each of the terms (5.2.3) is of equal magnitude. This minimizes the total entropy produced for a fixed number of cells. The effect of doubling the number of cells will be to double the number of such terms, while reducing their individual magnitudes by a factor of four. The net effect is to cut the entropy production in half. For this problem, the entropy production rate is inversely proportional to the number of grid points, other things being equal. This is characteristic of a first-order accurate scheme.

This is a first-order upwind scheme. It is well known in the literature and is highly dissipative unless an entropy destroying time advance scheme is used. For the special case of explicit Euler time advance with time step $\Delta t = \frac{\Delta x}{c}$, the errors in the time advance scheme precisely cancel the errors in the space differencing scheme and no entropy is produced. At larger values of the time step, the time differencing errors dominate and a net destruction of entropy results. The scheme is then unstable.

5.3 A Second-Order Scheme

Construction of second-order accurate schemes which satisfy a cell entropy inequality was discussed in Section 3.5. In this section the wave equation is used to illustrate the technique. The semi-discrete formula for entropy production (5.1.8) is restated for convenience.

$$\Delta x(\dot{P}_s)_j = c(u_j - u_{j-\frac{1}{2}})^2 - c(u_{j+\frac{1}{2}} - u_j)^2 \geq 0 \tag{5.3.1}$$

Assume $c > 0$. A useful strategy is to reduce the magnitude of the negative (second) term. If nothing is known about the magnitude of the first term, then the second term has to be reduced to zero to assure satisfaction of (5.3.1). That strategy leads to the first-order upwind scheme explored in the previous section. To improve on it, consider schemes in which

$$|u_j - u_{j-\frac{1}{2}}| \geq \frac{1}{2}|u_j - u_{j-1}| \tag{5.3.2}$$

This convention establishes a (usually) nonzero lower bound on the magnitude of the first term in (5.3.1). The magnitude of the second term need only be reduced to $\frac{1}{2}|u_j - u_{j-1}|$. Given the constraint (5.3.2), equation (5.3.1) reduces to

$$|u_{j+\frac{1}{2}} - u_j| \leq \frac{1}{2}|u_j - u_{j-1}| \tag{5.3.3}$$

The second-order accurate linear approximation

$$u_{j+\frac{1}{2}} = \frac{1}{2}(u_{j+1} + u_j) \tag{5.3.4}$$

may satisfy (5.3.3). If it doesn't, a different second-order approximation can be used instead

$$u_{j+\frac{1}{2}} = \frac{3}{2}u_j - \frac{1}{2}u_{j-1} \tag{5.3.5}$$

The two formulæ are equivalent if $(u_{j+1} - u_j) = (u_j - u_{j-1})$. This can be summarized as follows

$$u_{j+\frac{1}{2}} = \begin{cases} u_j + \frac{1}{2}(u_{j+1} - u_j) & \text{if} \quad |u_{j+1} - u_j| \leq |u_j - u_{j-1}| \\ u_j + \frac{1}{2}\left|\frac{u_j - u_{j-1}}{u_{j+1} - u_j}\right|(u_{j+1} - u_j) & \text{otherwise} \end{cases} \tag{5.3.6}$$

As in the first-order scheme, no conflicts arise. If $|u_{j+\frac{1}{2}} - u_j|$ is reduced, $|u_{j+1} - u_{j+\frac{1}{2}}|$ will increase, allowing (5.3.2) to be satisfied at grid point $j+1$.

The commonly used MINMOD limiter differs from (5.3.6) in its use of the first-order formula $u_{j+\frac{1}{2}} = u_j$ at extrema. At extrema, (5.3.6) behaves slightly better than the MINMOD limiter. The first reference to (5.3.6) is in an earlier publication by the author (ref. 25).

As in the first-order scheme, the total rate of entropy production can be computed. This computation requires understanding the switch given in (5.3.6). If $u_{j+\frac{1}{2}} = \frac{1}{2}(u_j + u_{j+1})$, the first branch, the scheme is exactly central differencing. The second branch evaluates to $u_{j+\frac{1}{2}} = \frac{3}{2}u_j - \frac{1}{2}u_{j-1}$ except at extrema. This corresponds to a second order upwind scheme of Warming and Beam.

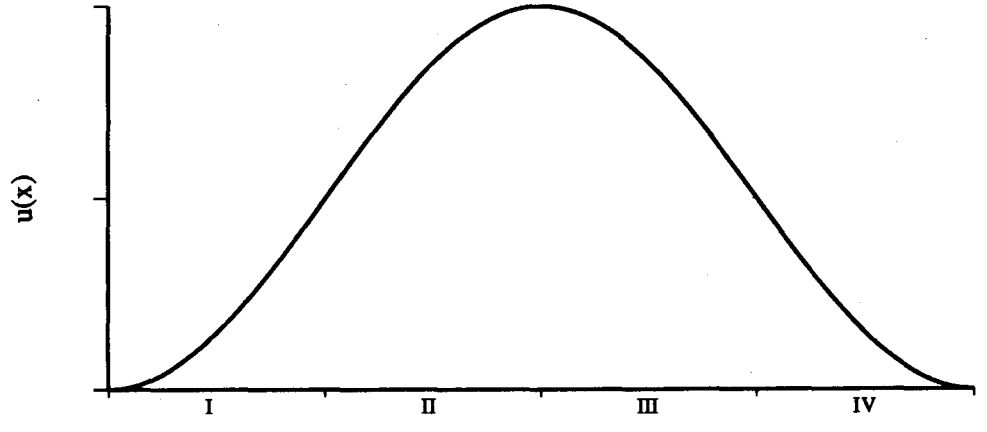

Figure 5.1. – A smooth pulse.

Consider the special case where $u(x)$ is the smooth pulse shown in figure 5.1. In regions I and III, a second-order upwind scheme is used. In these instances, the quantity $(u_{j+\frac{1}{2}} - u_j)$, which appears in the entropy production rate, is given by

$$(u_{j+\frac{1}{2}} - u_j) = \frac{1}{2}(u_j - u_{j-1}) \tag{5.3.7}$$

and the quantity $(u_j - u_{j-\frac{1}{2}})$, which also contributes to the entropy production is given by

$$(u_j - u_{j-\frac{1}{2}}) = \frac{1}{2}(2u_j - 3u_{j-1} + u_{j-2}) \tag{5.3.8}$$

When the entropy production is computed according to (5.3.1), using formulas (5.3.7) and (5.3.8), the result factors neatly

$$\Delta x(\dot{P}_s)_j = \frac{c}{4}(3u_j - 4u_{j-1} + u_{j-2})(u_{j-2} - 2u_{j-1} + u_j) \tag{5.3.9}$$

The first factor is a first derivative formula (except for the Δx) and the second factor is a second derivative formula (without the Δx^2). In fact, a Taylor series shows (dividing out the Δx)

$$(\dot{P}_s)_j = \frac{c}{2}u_{,x}u_{,xx}\Delta x^2 + O(\Delta x^3) \tag{5.3.10}$$

In regions II and IV, a second-order central differencing scheme is used. In these regions, the quantity $(u_{j+\frac{1}{2}} - u_j)$, which appears in the entropy production rate, is given by

$$(u_{j+\frac{1}{2}} - u_j) = \frac{1}{2}(u_{j+1} - u_j) \tag{5.3.11}$$

and the quantity $(u_j - u_{j-\frac{1}{2}})$, which also appears is given by

$$(u_j - u_{j-\frac{1}{2}}) = \frac{1}{2}(u_j - u_{j-1}) \tag{5.3.12}$$

When the entropy production is computed according to (5.3.1), using formulas (5.3.11) and (5.3.12), the result again factors neatly

$$\Delta x(\dot{P}_s)_j = -\frac{c}{4}(u_{j+1} - u_{j-1})(u_{j-1} - 2u_j + u_{j+1}) \tag{5.3.13}$$

The first factor is a first derivative formula (except for the Δx) and the second factor is a second derivative formula (without the Δx^2). In fact, a Taylor series expansion shows

$$(\dot{P}_s)_j = -\frac{c}{2}u_{,x}u_{,xx}\Delta x^2 + O(\Delta x^3) \tag{5.3.14}$$

In equations (5.3.10) and (5.3.14) $u_{,x}$ and $u_{,xx}$ are evaluated at point j, for which the entropy production is being calculated. Notice that, in both cases, the entropy production rate declines as the square of the mesh size. This implies second-order spatial accuracy.

Notice that (5.3.10) and (5.3.14) are identical except for sign (and higher order terms). This is because the product $u_{,x}u_{,xx}$ is positive in regions I and III while it is negative in regions II and IV. Thus the entropy production rate $(\dot{P}_s)_j$ is positive in either case. Incidentally, equations (5.3.9) and (5.3.13) both evaluate exactly to zero for linear data. This is another test for second-order accuracy.

The remaining cases to consider are the transitions. There are three different kinds. Between regions I and II or between regions III and IV, $(u_{j+\frac{1}{2}} - u_j)$ is given by (5.3.11)

while $(u_j - u_{j-\frac{1}{2}})$ is given by (5.3.8). The product, $(u_{,x}u_{,xx})$, goes through zero here because $u_{,xx}$ goes through zero. The entropy production rate factors to

$$\Delta x(\dot{P}_s)_j = (u_{j+1} - u_j - 3u_{j-1} + u_{j-2})(-u_{j+1} + 3u_j - 3u_{j-1} + u_{j-2}) \tag{5.3.15}$$

The first factor is a first derivative formula while the second factor is a third derivative formula. The Taylor series expansion shows

$$(\dot{P}_s)_j = -\frac{1}{2}cu_{,x}u_{,xxx}\Delta x^3 + O(\Delta x^4) \tag{5.3.16}$$

This shows that such transitions (inflection points) produce entropy at a lower rate than the basic scheme.

Between regions II and III, $u(x)$ reaches a maximum and $u_{,x} = 0$. The computation of $u_{j+\frac{1}{2}}$ may involve either branch of (5.3.6), depending on the relative magnitudes of the intervals on either side of the peak.

If the interval to the left of the peak is larger in magnitude than the interval to the right of the peak, the transition point will be the first point to the right of the peak. At the transition point (j), $u_{j+\frac{1}{2}}$ is computed by the usual upwind formula (5.3.5) and $(u_{j+\frac{1}{2}} - u_j)$ is given by (5.3.7). On the other hand, $u_{j-\frac{1}{2}}$ is computed by the symmetric formula

$$u_{j-\frac{1}{2}} = \frac{1}{2}(u_j + u_{j-1}) \tag{5.3.17}$$

and $(u_j - u_{j-\frac{1}{2}})$ is given by (5.3.12). Notice that $(u_{j+\frac{1}{2}} - u_j) = (u_j - u_{j-\frac{1}{2}})$. Consequently, no entropy is produced by such transitions.

If the interval to the left of the peak is smaller in magnitude than the interval to the right of the peak, the transition point will be the peak itself. At the transition point (j), $u_{j+\frac{1}{2}}$ is computed by the second branch of (5.3.6). The absolute value sign is needed in this case and $u_{j+\frac{1}{2}}$ is given by

$$u_{j+\frac{1}{2}} = \frac{1}{2}(u_j + u_{j-1}) \tag{5.3.18}$$

A simple manipulation gives

$$(u_{j+\frac{1}{2}} - u_j) = -\frac{1}{2}(u_j - u_{j-1}) \tag{5.3.19}$$

On the other hand, $u_{j-\frac{1}{2}}$ is computed by the symmetric formula (5.3.17) and, $(u_j - u_{j-\frac{1}{2}})$ is given by (5.3.12). Notice that $(u_{j+\frac{1}{2}} - u_j) = -(u_j - u_{j-\frac{1}{2}})$. Since these differences are squared in (5.3.1), no entropy is produced by these transitions either.

To summarize, the scheme defined by (5.3.6) is second-order accurate from the point of view of entropy production rate errors. There is no entropy production at extrema (a property not shared by the standard MINMOD limiter). More importantly, the entropy production rate for this scheme is nonnegative on a cell-by-cell basis.

5.4 An Explicit Scheme Which Satisfies a Cell Entropy Inequality

If the first-order upwind differencing scheme of Section 5.2 is advanced in time using explicit Euler time advance, the result satisfies the fully discrete entropy inequality for the wave equation provided $\nu \equiv \frac{c\Delta t}{\Delta x} \leq 1$. This well-known result can be shown by direct substitution. The semi-discrete scheme is just

$$(u_{,t})_j + \frac{c}{\Delta x}(u_j - u_{j-1}) = 0 \tag{5.4.1}$$

This leads to the fully discrete scheme, which, after multiplying by Δt is

$$u_j^{n+1} - u_j^n + \nu(u_j^n - u_{j-1}^n) = 0 \tag{5.4.2}$$

The fully discrete entropy inequality is

$$(\dot{P}_s)_j = \frac{S_j^{n+1} - S_j^n}{\Delta t} + \frac{F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n}{\Delta x} \tag{5.4.3}$$

Substituting the definitions (5.1.6) gives

$$\Delta t(\dot{P}_s)_j = -\left[(u_j^{n+1})^2 - (u_j^n)^2\right] - \nu\left[(u_j^n)^2 - (u_{j-1}^n)^2\right] \tag{5.4.4}$$

The only part of (5.4.4) that is not directly computable is $(u_j^{n+1})^2$ and this can be found by solving (5.4.2) for u_j^{n+1} and substituting. When this is done the expression for entropy production rate reduces to

$$\Delta t(\dot{P}_s)_j = \nu(1-\nu)(u_j^n - u_{j-1}^n)^2 \tag{5.4.5}$$

This is a positive quantity only if $0 \leq \nu \leq 1$. This corresponds to the von Neumann stability range of the scheme.

It is possible to produce a more accurate scheme if a more restrictive time step range can be accepted. In general the production of entropy in a single step for the wave equation using explicit Euler time advance is given by

$$\Delta t(\dot{P}_s)_j = \nu\left[-(u_{j+\frac{1}{2}}^n - u_j^n)^2 + (u_j^n - u_{j-\frac{1}{2}}^n)^2\right] - \nu^2(u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n)^2 \tag{5.4.6}$$

which is derived from (4.2.8) and (5.1.8). The right-hand side must be positive to satisfy the second law. The last term makes this difficult, especially if ν is large. Equation (5.4.6) can be restated as

$$\Delta t(\dot{P}_s)_j = \nu(u_j^n - u_{j-\frac{1}{2}}^n)^2\left[\big(R+1\big)\big((1-\nu) - R(1+\nu)\big)\right] \tag{5.4.7}$$

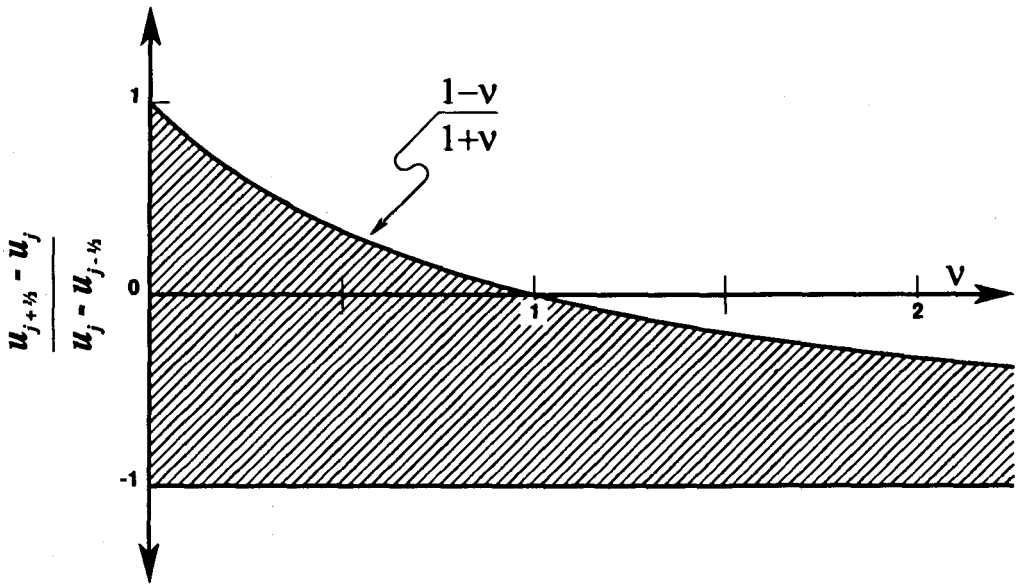

Figure 5.2. – Only schemes in the shaded region satisfy a cell entropy inequality.

where

$$R \equiv \frac{u^n_{j+\frac{1}{2}} - u^n_j}{u^n_j - u^n_{j-\frac{1}{2}}} \tag{5.4.8}$$

Assuming $\nu > 0$ (the case where $\nu < 0$ is a trivial extension), it is clear by inspection of (5.4.7) that $(\dot{P}_s)_j > 0$ when

$$-1 \leq R \leq \frac{1-\nu}{1+\nu} \tag{5.4.9}$$

The two inequalities are satisfied within the shaded region of figure 5.2.

Notice that for CFL numbers (ν) greater than one, only negative values of the ratio can be admitted. If $u^n_j = u^n_{j+1}$ then consistency suggests that $u^n_{j+\frac{1}{2}} = u^n_j$. Referring to (5.4.8) this implies that $R = 0$. Thus, it is not generally possible to construct schemes having only negative values of R.

At a CFL number of 1, the (5.4.9) reduces to $u_{j+\frac{1}{2}} = u_j$, which is described above. For CFL numbers less than or equal to one, it is possible to construct consistent schemes which satisfy a cell entropy inequality. Since explicit Euler destroys entropy, the second-order scheme (5.3.6) described in Section 5.3 must be modified to produce more entropy.

This can be done as follows:

$$u_{j+\frac{1}{2}} = \begin{cases} u_j + \frac{1}{2}(1-\nu)(u_{j+1}-u_j) & \text{if} \quad |u_{j+1}-u_j| \le |u_j - u_{j-1}| \\ u_j + \frac{1}{2}(1-\nu)\left|\frac{u_j - u_{j-1}}{u_{j+1}-u_j}\right|(u_{j+1}-u_j) & \text{otherwise} \end{cases} \tag{5.4.7}$$

The modified scheme which results will not be second-order accurate unless Δt is of the same order as Δx^2. In that case $(1-\nu)$ is essentially equal to one and the scheme reverts to one previously analyzed (5.3.6). Unfortunately, such a time step restriction is generally too severe. Nevertheless, (5.4.7) is much more accurate than (5.2.1) for any stable value of ν.

For transient problems such as this, the spatial and time discretizations are inseparably coupled. Any entropy produced by either will stay until it convects through an outflow boundary or until the calculation stops. It can never be removed by a scheme which satisfies a cell entropy inequality.

5.5 A Stable TVD Scheme Which Violates the Cell Entropy Condition

To construct a scheme and analyze it for compliance with the second law requires two things, a time advance scheme and a space differencing scheme. Space differencing schemes can be expressed as algorithms for computing $u_{j+\frac{1}{2}}$. Many such schemes can be characterized by the formula

$$u_{j+\frac{1}{2}} = u_j + \frac{1}{2}\phi_{j+\frac{1}{2}}(u_{j+1}-u_j) \tag{5.5.1}$$

This requires an algorithm for computing $\phi_{j+\frac{1}{2}}$. Usually $\phi_{j+\frac{1}{2}}$ is constrained to fall between 0 and 2, which is equivalent to constraining $u_{j+\frac{1}{2}}$ to lie between u_j and u_{j+1}. The first-order upwind scheme described above can be characterized by the formula $\phi_{j+\frac{1}{2}} = 0$, while central differencing is characterized (for the wave equation) by $\phi_{j+\frac{1}{2}} = 1$. Somewhat more subtle is the description of a second-order upwind scheme for which

$$\phi_{j+\frac{1}{2}} = \frac{u_j - u_{j-1}}{u_{j+1}-u_j} \tag{5.5.2}$$

The right-hand side is often shortened to r_j. This suggests that more complicated schemes might be represented as a plot of $\phi_{j+\frac{1}{2}}(r_j)$. Several schemes which can be characterized in this way are plotted in figure 5.3.

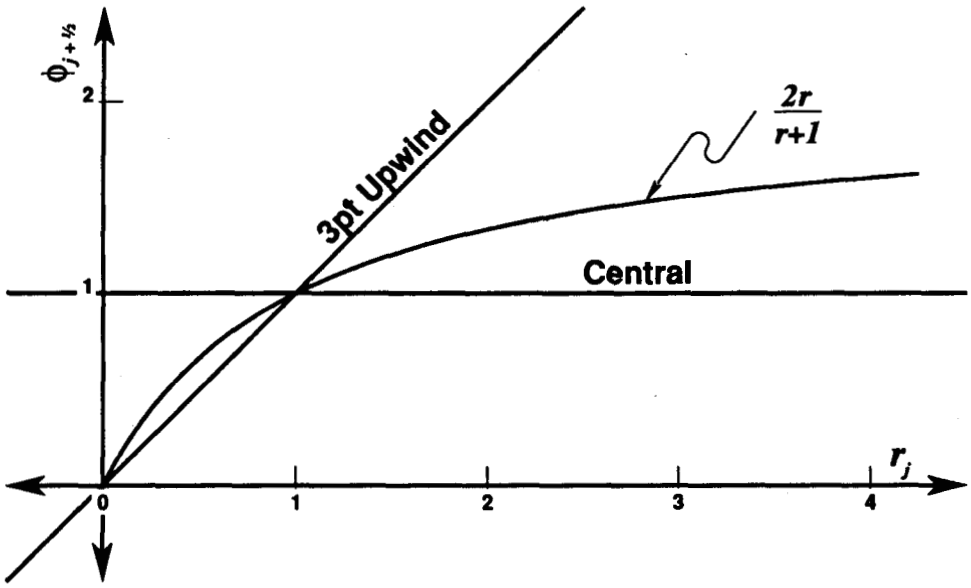

Figure 5.3. – Many common difference schemes can be defined in terms of $\phi(r)$.

A simple scheme which is stable, but violates the entropy condition, is

$$\phi_{j+\frac{1}{2}} = \begin{cases} 0 & \text{if } r_j < 0 \\ \dfrac{2r_j}{r_j+1} & \text{if } r_j \geq 0 \end{cases} \tag{5.5.3}$$

The corresponding semi-discrete scheme is spatially second-order accurate. When used with explicit Euler time advance, (5.5.3) is not only stable (conditionally), but Total Variation Diminishing (TVD) (ref. 25). There it was also shown analytically that the scheme violates entropy. The next section will show results obtained with this scheme.

5.6 Results

This section shows the results of computational experiments made with schemes described above. Two initial waveforms were used. One was the square pulse of unit height and a width of 10 grid points. The other was a single period of a sine wave with the same amplitude and width. To compare schemes with different stability limits, and to stay away from the perfect shift that occurs for some schemes when $\nu = 1$, all schemes were implemented with $\nu = \frac{1}{2}$. The results are shown after 50 steps. During that time the pulse has propagated 25 grid points.

The first scheme shown is the entropy violating scheme of Section 5.5, defined by (5.5.3). Remember that this scheme is spatially second-order accurate, conservative, stable, and TVD. What can possibly go wrong? In figure 5.4 we see that it does an excellent job of

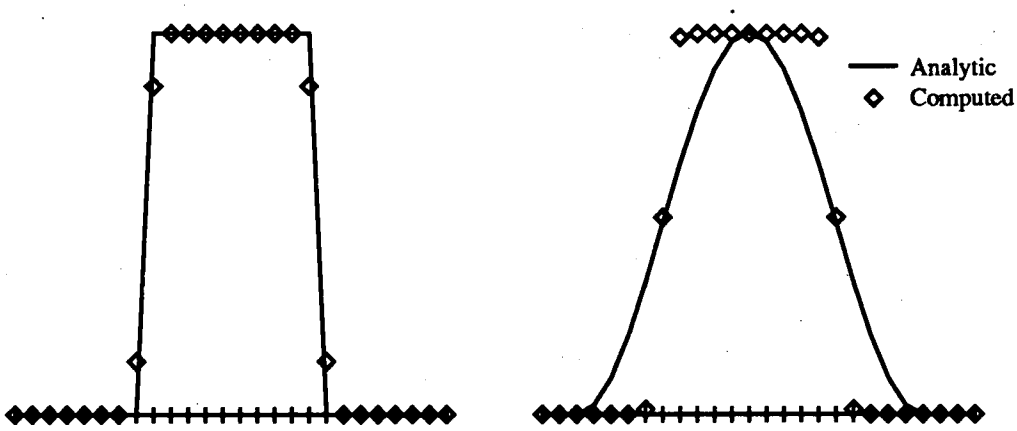

Figure 5.4. – This scheme satisfies a TVD condition, but does not satisfy a cell entropy inequality. Both pulses have traveled 2.5 pulse widths. Notice the tendency to produce a square wave from smooth initial data.

convecting a square wave. There is no noticeable amplitude loss, and the shape is virtually unchanged. The area under the pulse is conserved.

The right half of figure 5.4 shows why the square wave is captured so accurately. Many initial pulses will eventually become square waves. This behavior reflects the decreasing entropy (increasing L_2 norm) within the domain. Because of the TVD condition, the L_1 norm is bounded. This apparent contradiction is resolved only when $u(x)$ is piecewise constant as it is for the square pulse. Such solutions have zero entropy production. The square pulse minimizes the total entropy in the domain, subject to the conservation law and the TVD constraint.

The second scheme is the simple first-order upwind scheme, (5.2.1). This scheme is first-order accurate in space and time, conservative, and satisfies a cell entropy inequality. In figure 5.5 we see a substantial loss of amplitude and smearing of sharp corners that this scheme is famous for.

The third scheme shown (5.4.7) also satisfies a cell entropy inequality. It is slightly less dissipative than the first-order upwind scheme, but remains first-order accurate in time due to explicit Euler time advance. This scheme is very similar to the limited Lax-Wendroff schemes seen in early TVD papers (ref. 11). Only the limiter is changed slightly as previously noted. In figure 5.6 we see that this scheme is a big improvement over the previous one, at least for this equation. This improvement reflects the fact that the entropy produced by the spatial differencing approximately cancels the entropy destroyed by the temporal differencing. This strategy is much less effective in systems that do not have a single, constant, wave speed.

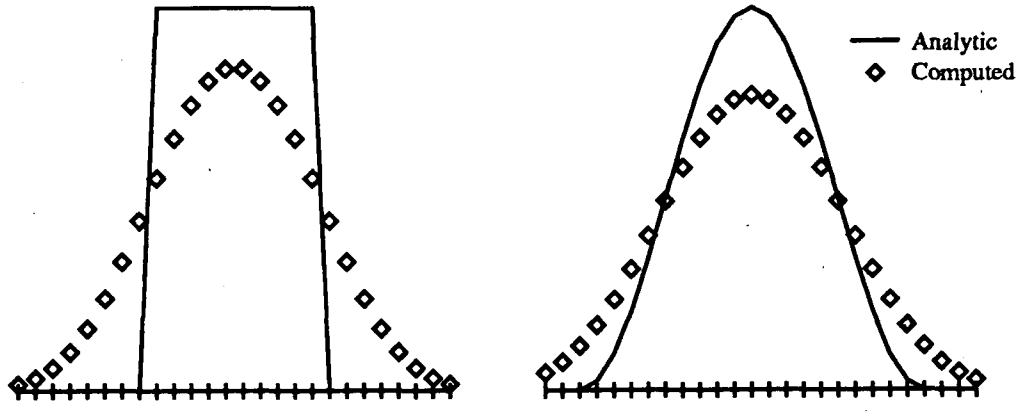

Figure 5.5. – These results are from a standard first-order upwind scheme which uses explicit Euler time advance. This scheme satisfies a cell entropy inequality.

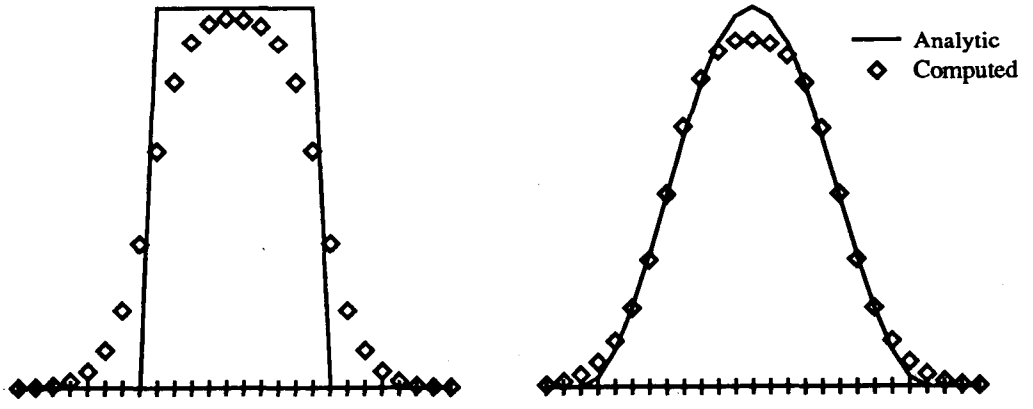

Figure 5.6. – This scheme is significantly more accurate than the standard first-order upwind scheme. It uses explicit Euler time advance and satisfies a cell entropy inequality. The accuracy improvement comes from a partial cancellation between space differencing and time differencing errors.

A scheme which is second-order accurate in space (5.3.6) was discussed in Section 5.3. This scheme satisfies a semi-discrete entropy inequality. Advancing in time using implicit Euler time advance preserves this inequality. Since this is a linear problem and since the limiter is piecewise differentiable, there is little difficulty in implementing this scheme. The results for this scheme are shown in figure 5.7.

The results are disappointing. The errors from the time advance dominate the calculation. The total entropy produced is not too different from that produced by a first-order upwind scheme. For problems with a steady state, the time advance errors would eventually disappear, leaving a second-order accurate solution at steady state.

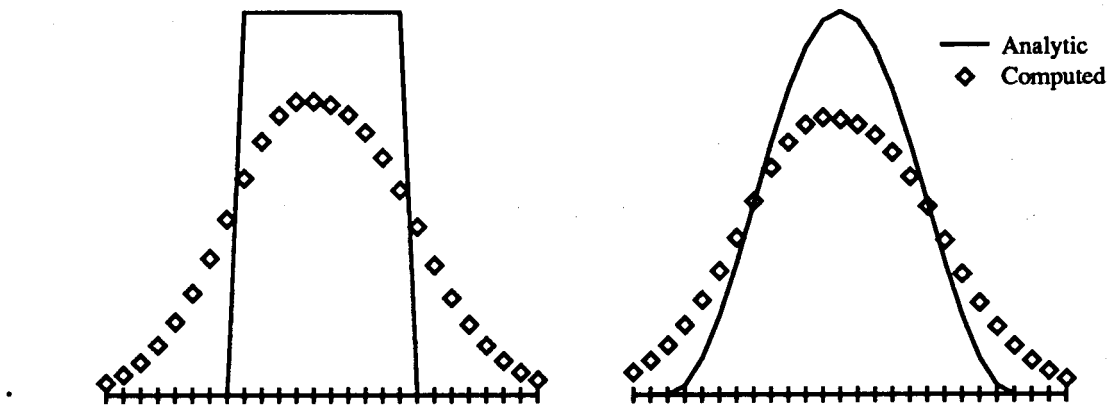

Figure 5.7. – This scheme is second-order accurate in space. First-order accurate, implicit Euler time advance is used. The time advance errors dominate.

To improve the time accuracy, a modified Crank–Nicolson time advance may be used. This was described in Chapter 4. Such a scheme eliminates the temporal dissipation in this case. The improvement can clearly be seen in figure 5.8.

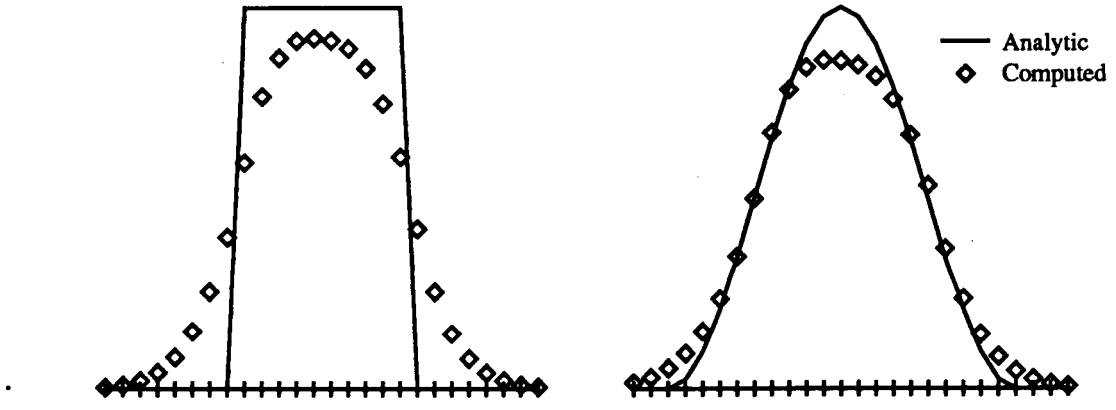

Figure 5.8. – This scheme uses a modified Crank–Nicolson time advance to achieve second-order time accuracy. Spatial differencing is the same as for figure 5.7.

A single number describing the accuracy of all the schemes is total entropy produced during the time of integration. This is easily computed, it is just the sum of u^2 over all grid points for the initial condition less the equivalent sum after 50 steps. For this simple problem, the correct answer is exactly zero. A table summarizing these results is given below

Table 5.1. – Entropy Production For Various Schemes

Figure	Square Pulse	Sine Pulse
5.4	0.48	-1.87
5.5	3.90	1.98
5.6	2.21	0.44
5.7	5.40	2.26
5.8	2.83	0.86

The clear winner is the first-order scheme of figure 5.6. Such a scheme is unsuited for problems having a steady state, since the time step affects the computed solution to such problems. Furthermore, the time step is limited to a CFL number of 1 in any event.

The Crank–Nicolson scheme of figure 5.8 has no such problems. It is formally second-order accurate in space and time. The spatial accuracy is independent of the time step, which is limited only by the nonlinearity of the problem (see Section 4.5). Most importantly, it satisfies a cell entropy inequality. In the next chapter, this scheme will be adapted to nonlinear problems.

Chapter 6. NONLINEAR SCALAR EQUATIONS

6.1 A Suitable Entropy for Burgers' Equation

This chapter is concerned with the solution of nonlinear scalar equations of the form

$$u_{,t} + f_{,x} = 0 \tag{6.1.1}$$

where f is any nonlinear, continuous function of u. In particular, Burgers' equation

$$u_{,t} + \frac{1}{2}(u^2)_{,x} = 0 \tag{6.1.2}$$

will be explored. As previously shown, a suitable entropy function for any scalar equation is

$$S = -u^2 \tag{6.1.3}$$

This satisfies the convexity condition.

$$S_{,uu} = -2 < 0 \tag{6.1.4}$$

For Burgers' equation, $f = \frac{1}{2}u^2$. The compatability condition

$$F_{,u} = S_{,u} f_{,u} \tag{6.1.5}$$

determines the corresponding entropy flux function F. This is

$$F = -\frac{2}{3}u^3 \tag{6.1.6}$$

as may be verified by substitution. Thus, a suitable entropy pair for Burgers' equation is given by equations (6.1.3) and (6.1.6).

6.2 A First-Order Scheme

Construction of first-order schemes which satisfy the second law was discussed in Section 3.4. In this section such a scheme is constructed for Burgers' equation.

In Chapter 4 it was shown that satisfaction of the semi-discrete entropy inequality implies satisfaction of the fully discrete inequality when certain time advance schemes are used. The semi-discrete form of (6.1.1) is

$$u_{,t} + \frac{\Delta t}{\Delta x}(f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}) \tag{6.2.1}$$

where $f_{j+\frac{1}{2}} \equiv f(u_{j+\frac{1}{2}})$. The choice of $u_{j+\frac{1}{2}}$ is subject to the constraint that the second law is satisfied in each cell. An expression for this constraint was derived in Section 3.4:

$$\begin{aligned}\Delta x_j(\dot{P}_s)_j = &\left[-(S_{,u})_j(f_{j+\frac{1}{2}} - f_j) + (F_{j+\frac{1}{2}} - F_j)\right] \\ &+ \left[-(S_{,u})_j(f_j - f_{j-\frac{1}{2}}) + (F_j - F_{j-\frac{1}{2}})\right] \geq 0 \end{aligned} \tag{6.2.2}$$

where $F_{j+\frac{1}{2}} \equiv F(u_{j+\frac{1}{2}})$. The terms in brackets define the right and left half cell entropy production rates, $\Delta x_j(\dot{P}_s^+)_j$ and $\Delta x_j(\dot{P}_s^-)_j$ respectively. Using these definitions (6.2.2) can be written as

$$\Delta x_j \left[(\dot{P}_s^+)_j + (\dot{P}_s^-)_j\right] \geq 0 \tag{6.2.3}$$

The algorithm proposed in Section 3.4 is to choose $u_{j+\frac{1}{2}}$ in such a way that $(\dot{P}_s^+)_j$ and $(\dot{P}_s^-)_j$ are individually positive. This requires that

$$(\dot{P}_s^+)_j \geq 0 \text{ and } (\dot{P}_s^-)_{j+1} \geq 0 \tag{6.2.4}$$

It is not immediately obvious that a value of $u_{j+\frac{1}{2}}$ exists which satisfies this requirement. The choices of $u_{j+\frac{1}{2}}$ may be parameterized in general by the formula

$$u_{j+\frac{1}{2}} = u_j + \frac{1}{2}\phi_{j+\frac{1}{2}}(u_{j+1} - u_j) \tag{6.2.5}$$

where the parameter $\phi_{j+\frac{1}{2}}$ takes on values between 0 and 2 as $u_{j+\frac{1}{2}}$ takes on values between u_j and u_{j+1}. If $0 \leq \phi_{j+\frac{1}{2}} \leq 2$, and $u_{j+\frac{1}{2}}$ is chosen according to equation (6.2.5) the following inequality holds.

$$(u_{j+1} - u_{j+\frac{1}{2}})(u_{j+\frac{1}{2}} - u_j) \geq 0 \tag{6.2.6}$$

This inequality is equivalent to (3.4.12) and allows the use of inequality (3.4.8) and equation (3.4.7).

Sometimes a single example is more illuminating than a page of algebra. In figure 6.1, the values of $\Delta x_j(\dot{P}_s^+)_j$ and $\Delta x_{j+1}(\dot{P}_s^-)_{j+1}$ are plotted as functions of $\phi_{j+\frac{1}{2}}$ for a strong compression, $u_j = 1$ and $u_{j+1} = \frac{1}{2}$.

When $\phi_{j+\frac{1}{2}} = 0$, $u_{j+\frac{1}{2}} = u_j$ and $(\dot{P}_s^+)_j = 0$ by inspection of equation (6.2.2). Similarly, when $\phi_{j+\frac{1}{2}} = 2$, $u_{j+\frac{1}{2}} = u_{j+1}$ and $(\dot{P}_s^-)_{j+1} = 0$. Figure 6.1 confirms this behavior for a particular case.

Also in figure 6.1, it is clear that, for any value of $\phi_{j+\frac{1}{2}}$ in the range plotted, the slopes of the two curves have the same sign. This is a trivial extension of inequality

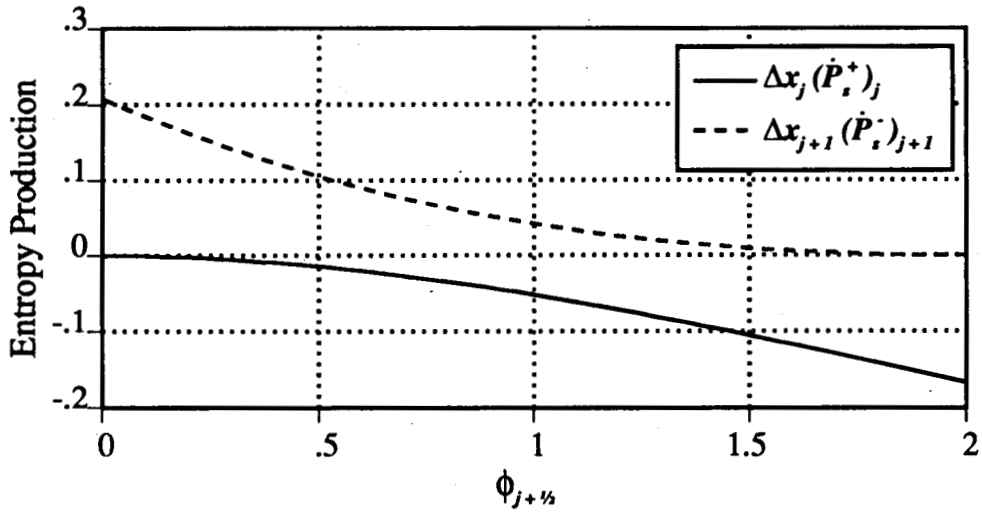

Figure 6.1. – Effect of the interpolation parameter ϕ on semi-discrete cell entropy production rates for Burgers' equation during a strong compression.

(3.4.12). Inequality (3.4.8) shows that the product of the two slopes will be positive if and only if (6.2.6) holds. This feature will be exploited more heavily in the next section.

In addition to having the same sign at a given value of $\phi_{j+\frac{1}{2}}$ the two curves have no extrema on the interval if u_j and u_{j+1} have the same sign. Extending (3.4.6) through the chain rule and simplifying gives the formula

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial \phi_{j+\frac{1}{2}}} = -\frac{1}{2}(u_{j+1} - u_j)^2 \phi_{j+\frac{1}{2}} u_{j+\frac{1}{2}} \tag{6.2.7}$$

which is valid for only Burgers' equation. The identity $S_{,uu} = -2$ was used in the simplification. Equation (6.2.7) shows that, throughout the range of $\phi_{j+\frac{1}{2}}$ shown, the slope of both curves is opposite in sign to the value of $u_{j+\frac{1}{2}}$. In this case $u_{j+\frac{1}{2}}$ is positive throughout the interval so both curves have negative slopes everywhere.

As long as u_j and u_{j+1} have the same sign, the choice

$$\phi_{j+\frac{1}{2}} = 1 - \text{sign}(u_j) \tag{6.2.8}$$

will always satisfy the inequalities of (6.2.4). These, in turn guarantee a cell entropy inequality.

If u_j and u_{j+1} have opposite signs, the two curves will contain extrema. In figure 6.2, for example, $u_j = -1$ and $u_{j+1} = 1$.

Notice that, despite the extrema, the two curves have slopes of the same sign for any given value of $\phi_{j+\frac{1}{2}}$. This implies that the extrema will be at the same point on the two

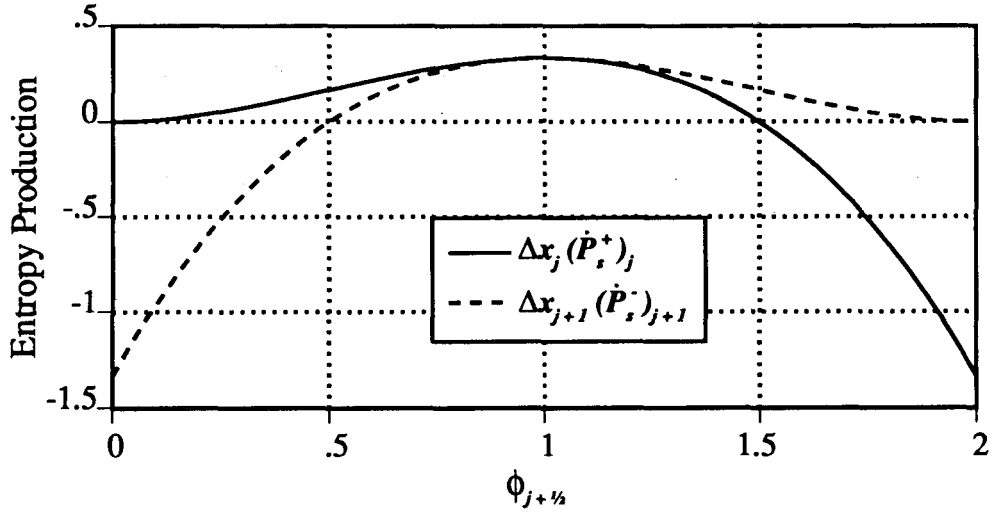

Figure 6.2. – During a sonic expansion, both cells experience a maximum entropy production rate when $u_{j+\frac{1}{2}} = 0$. In this case that occurs when $\phi_{j+\frac{1}{2}} = 1$.

curves. This is a case where neither of the endpoints satisfy (6.2.4). One point which will is the extremum; $\phi_{j+\frac{1}{2}} = 1$ in this case. For sonic points such as this, the extremum occurs when

$$(f_{,u})_{j+\frac{1}{2}} = 0 \tag{6.2.9}$$

It is only at sonic points such as this that the first-order scheme given here differs from the usual first-order upwind scheme. Incidentally, these are called sonic points by analogy to the one-dimensional Euler equations. They are identified by a change in sign of the wave speed in a region where the characteristics diverge.

For Burgers' equation, (6.2.9) is equivalent to $u_{j+\frac{1}{2}} = 0$. Often, it is easier to implement the formula

$$\phi_{j+\frac{1}{2}} = -\frac{2u_j}{u_{j+1} - u_j} \tag{6.2.10}$$

because it specifies $\phi_{j+\frac{1}{2}}$ instead of $u_{j+\frac{1}{2}}$. Substituting (6.2.10) into (6.2.5) yields $u_{j+\frac{1}{2}} = 0$.

In figure 6.2, which depicts a sonic point, the two curves contain maxima. In the case of a shock, the extrema are minima. Figure 6.3 depicts such a case, with $u_j = 1$ and $u_{j+1} = -0.45$.

The minimum for both curves is at $\phi_{j+\frac{1}{2}} \approx 1.4$. Because they each have a single extremum, both curves will reach their maximum values at one of the endpoints. One choice which always satisfies (6.2.4) is

$$\phi_{j+\frac{1}{2}} = 1 - \text{sign}\left(\frac{u_j + u_{j+1}}{2}\right) \tag{6.2.11}$$

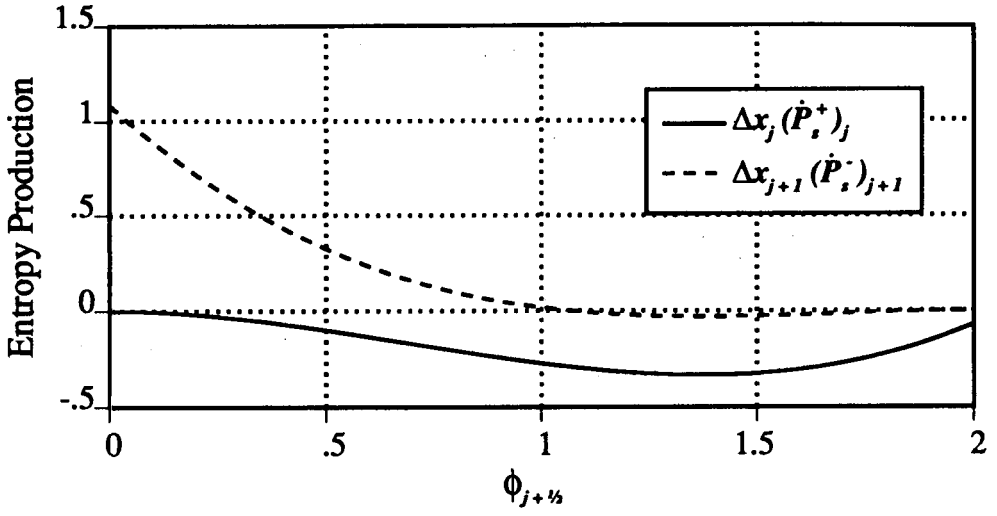

Figure 6.3. – A rapid compression through a moving shock.

This evaluates to one of the endpoints, $\phi_{j+\frac{1}{2}} = 0$ for this example. In cases where the shock is nearly stationary, roundoff error may cause the wrong endpoint to be selected. Fortunately, in such cases, either endpoint will satisfy (6.2.4), so this is not a problem.

In summary, a first-order scheme which satisfies a cell entropy inequality is defined by

$$\phi_{j+\frac{1}{2}} = \begin{cases} -\frac{2u_j}{u_{j+1}-u_j} & \text{sonic points } (u_j \le 0 < u_{j+1}) \\ 1 - \text{sign}\left(\frac{u_j+u_{j+1}}{2}\right) & \text{shocks } (u_j \ge 0 > u_{j+1}) \\ 1 - \text{sign}(u_j) & \text{elsewhere } (u_j u_{j+1} > 0) \end{cases} \tag{6.2.12}$$

The resulting scheme is exactly equivalent to a standard first-order upwind scheme except at shocks and sonic points. It is precisely at these points that entropy production is important. In the case of sonic points, entropy must not be destroyed in a numerical expansion shock. In the case of shocks, entropy must be produced as required by physics. Any numerical scheme without these properties will generate qualitatively incorrect solutions.

6.3 A Second-Order Scheme

The scheme devised in Section 6.2 is consistent and stable, but only first order accurate. Substantially more entropy is produced than is required by the physics. This section explores a variation which is second-order accurate. The technique was outlined in Chapter 3 and demonstrated in Chapter 5 on the linear wave equation.

A scheme which is consistent and second-order accurate is

$$\phi_{j+\frac{1}{2}} = 1 \tag{6.3.1}$$

This is not quite the same as central differencing because $f(u)$ is nonlinear. Although this scheme does not (in general) satisfy a cell entropy inequality, it may be possible to construct a scheme which does, in such a way as to approach (6.3.1) in the limit as the grid is refined. Such a scheme would be second order accurate in this limit.

The cell entropy inequality for the j^{th} cell is given by

$$(\dot{P}_s)_j = (\dot{P}_s^+)_j + (\dot{P}_s^-)_j \geq 0 \tag{6.3.2}$$

Where the first-order scheme of Section 6.2 required that these terms be individually positive, the second-order scheme requires only that their sum be positive in each cell.

The entropy production rate for the scheme defined by (6.3.1) might be written

$$(\dot{P}_s)_j = \overline{(\dot{P}_s^+)_j} + \overline{(\dot{P}_s^-)_j} \tag{6.3.3}$$

where $\overline{(\dot{P}_s^+)_j}$ is the value of $(\dot{P}_s^+)_j$ which corresponds to $u_{j+\frac{1}{2}} = \frac{1}{2}(u_{j+1} + u_j)$. The entropy production rate $(\dot{P}_s)_j$ might be positive or negative, depending on the data.

Now consider schemes for which

$$(\dot{P}_s^+)_j \geq \overline{(\dot{P}_s^+)_j} \quad \text{and} \quad (\dot{P}_s^-)_j \geq \overline{(\dot{P}_s^-)_j} \tag{6.3.4}$$

Under this convention the inequality

$$(\dot{P}_s^+)_j + (\dot{P}_s^-)_j \geq (\dot{P}_s^+)_j + \overline{(\dot{P}_s^-)_j} \tag{6.3.5}$$

holds. If the right-hand side is positive, the left-hand side is also positive and the entropy inequality (6.3.2) is satisfied. This is helpful since the right-hand side of (6.3.5) has only a single free parameter, $\phi_{j+\frac{1}{2}}$, which determines $u_{j+\frac{1}{2}}$ (and ultimately $(\dot{P}_s^+)_j$). Unfortunately, satisfaction of (6.3.5) may not be possible; for example, if $u_j = u_{j+1}$ and $\overline{(\dot{P}_s^-)_j} < 0$. This necessitates modifying (6.3.4) as follows:

$$(\dot{P}_s^+)_j \geq \min\left(-\overline{(\dot{P}_s^-)_j}, 0\right) \tag{6.3.6}$$

$$(\dot{P}_s^-)_j \geq \min\left(-\overline{(\dot{P}_s^+)_j}, 0\right) \tag{6.3.7}$$

In the last section it was shown that $\phi_{j+\frac{1}{2}}$ can always be chosen in such a way as to make $(\dot{P}_s^+)_j$ and $(\dot{P}_s^-)_{j+1}$ nonnegative. This is sufficient to show that relations (6.3.6) and (6.3.7) can always be satisfied. In combination with (6.3.4), these are sufficient to satisfy a cell entropy inequality.

Since both (6.3.6) and (6.3.7) must be satisfied for each cell, the choice of $\phi_{j+\frac{1}{2}}$ must satisfy an additional inequality which is just (6.3.7) shifted by one index

$$(\dot{P}_s^-)_{j+1} \geq \min\left(-\overline{(\dot{P}_s^+)_{j+1}}, 0\right) \tag{6.3.8}$$

This is illustrated graphically in figure 6.4. As in figure 6.1, $u_j = 1$, and $u_{j+1} = 0.5$. Since $(\dot{P}_s^-)_{j+1} \geq 0$, condition (6.3.8) will automatically be satisfied. The horizontal chaindashed line indicates the value of $-\overline{(\dot{P}_s^-)_j}$. This value depends on u_j and u_{j-1} ($u_{j-1} = 1.25$ in this case) and is solution dependent. From the figure, one can see that a value of $\phi_{j+\frac{1}{2}} \approx 0.75$ would be chosen. This satisfies the inequalities (6.3.4), (6.3.6), and (6.3.8) and thus meets the cell entropy inequality for the j^{th} cell. If the value of $-\overline{(\dot{P}_s^-)_j}$ were lower (perhaps -0.1), a value of $\phi_{j+\frac{1}{2}} = 1$ would suffice. This is a much less dissipative method than the first-order scheme of Section 6.2. That scheme would have chosen the value $\phi_{j+\frac{1}{2}} = 0$.

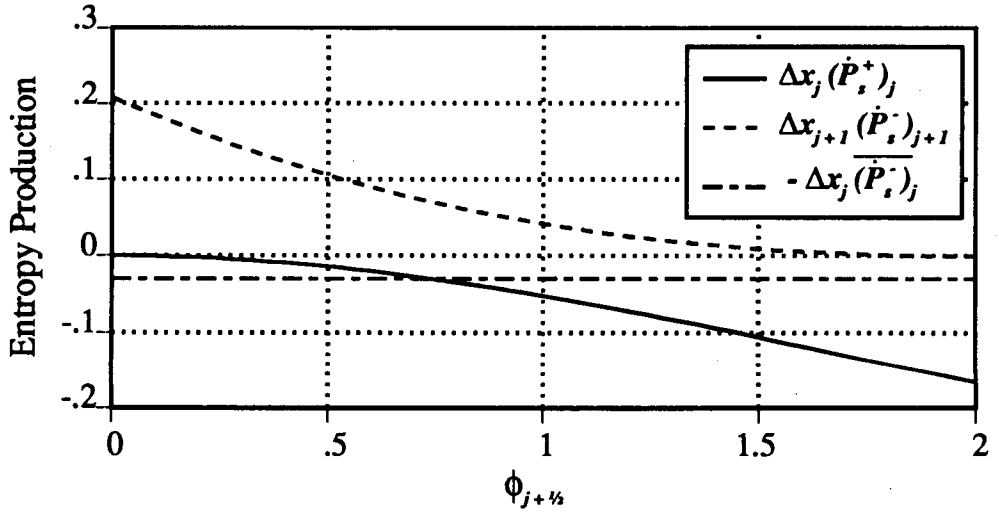

Figure 6.4. – Lowering the value of ϕ to 0.75 guarantees a positive value of $(\dot{P}_s)_j$. The value of $(\dot{P}_s)_{j+1}$ is also increased slightly, a side effect of the technique.

6.4 Practical Considerations

The scheme described in Section 6.3 requires computing the intersection point between a horizontal line and a curve. This was done graphically in figure 6.4; in practice a modified Newton's method can be used. There is also an elegant approximation that works well for Burgers' equation.

The Newton's method approach uses the expression for $\Delta x_j(\dot{P}_s^+)_j$ which is, from (6.2.2)

$$\Delta x_j(\dot{P}_s^+)_j \equiv -(S_{,u})_j(f_{j+\frac{1}{2}} - f_j) + (F_{j+\frac{1}{2}} - F_j) \tag{6.4.1}$$

Using the expressions given in Section 6.1 for $u_{j+\frac{1}{2}}$ this becomes, for Burgers' equation

$$\Delta x_j(\dot{P}_s^+)_j \equiv 2u_j\left(\frac{1}{2}{u_{j+\frac{1}{2}}}^2 - \frac{1}{2}{u_j}^2\right) + \left(-\frac{2}{3}{u_{j+\frac{1}{2}}}^3 + \frac{2}{3}{u_j}^3\right) \tag{6.4.2}$$

This expression factors

$$\Delta x_j(\dot{P}_s^+)_j = -\left(u_{j+\frac{1}{2}} - u_j\right)^2\left(u_j + \frac{2}{3}(u_{j+\frac{1}{2}} - u_j)\right) \tag{6.4.3}$$

Equation (6.2.5) expresses $u_{j+\frac{1}{2}}$ in terms of u_{j+1} and $\phi_{j+\frac{1}{2}}$. This allows computation of the gradient which is required for Newton's method.

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial\phi_{j+\frac{1}{2}}} = -\frac{1}{2}(u_{j+1} - u_j)^2\phi_{j+\frac{1}{2}}u_{j+\frac{1}{2}} \tag{6.4.4}$$

This expression for the gradient has three zeroes which must be guarded if Newton's method is to be reliable. In the first case, if $u_{j+1} - u_j = 0$, the choice of $\phi_{j+\frac{1}{2}}$ is arbitrary and has no effect on the choice of $u_{j+\frac{1}{2}}$. In the second case, $\phi_{j+\frac{1}{2}} = 0$, which can occur during strong compressions or shocks. The third case, $u_{j+\frac{1}{2}} = 0$, occurs at sonic points and shocks. The value of $\phi_{j+\frac{1}{2}}$ at which this occurs is

$$\phi_{j+\frac{1}{2}} = -\frac{2u_j}{u_{j+1} - u_j} \tag{6.4.5}$$

In the first case, the choice $\phi_{j+\frac{1}{2}} = 1$ is made. In the second and third cases, the change suggested by one step of Newton's method is limited to 90% of the distance to the nearest extrema. Finally, the magnitude of the change is limited to keep $\phi_{j+\frac{1}{2}}$ between 0 and 2, in case the initial guess is very close to an extremum.

In practice, this works quite well and converges in about four iterations to six decimal places. It works equally well for $(\dot{P}_s^-)_j$ for which the expression is

$$\Delta x_j(\dot{P}_s^-)_j = \left(u_j - u_{j-\frac{1}{2}}\right)^2\left(u_j - \frac{2}{3}(u_j - u_{j-\frac{1}{2}})\right) \tag{6.4.6}$$

and for which the gradient is

$$\Delta x_j \frac{\partial(\dot{P}_s^-)_j}{\partial\phi_{j-\frac{1}{2}}} = \frac{1}{2}(u_j - u_{j-1})^2(2 - \phi_{j-\frac{1}{2}})u_{j-\frac{1}{2}} \tag{6.4.7}$$

At sonic points, it is possible that both $\overline{(\dot{P}_s^+)_j}$ and $\overline{(\dot{P}_s^-)_j}$ are negative. This can result in the null stencil, $f_{j+\frac{1}{2}} = f_{j-\frac{1}{2}} = f_j = 0$, which implies $(u_{,t})_j = 0$. This only occurs if the sonic point happens to lie on a grid point, in which case $(u_{,t})_j$ is supposed to be zero.

At shocks $(\dot{P}_s^+)_{j-1}$ and $(\dot{P}_s^-)_j$ may both be negative. This results in two different solutions for $\phi_{j-\frac{1}{2}}$. The value farthest from 1 is selected naturally by Newton's method.

An approximation is often made to be computationally efficient. Referring to equation (6.4.3), if $\frac{2}{3}(u_{j+\frac{1}{2}} - u_j)$ is small compared to u_j, it can be neglected. In that case (6.4.3) reduces to

$$\Delta x_j(\dot{P}_s^+)_j \approx -\frac{1}{2}\phi_{j+\frac{1}{2}} u_j (u_{j+1} - u_j)^2 \tag{6.4.8}$$

Equation (6.4.8) can be solved with only a square root, but even this can be avoided. If $(\dot{P}_s^-)_j$ is approximated in the same way as $(\dot{P}_s^+)_j$, the result is

$$\Delta x_j(\dot{P}_s^-)_j \approx u_j(\frac{1}{2}(2 - \phi_{j-\frac{1}{2}})(u_j - u_{j-1}))^2 \tag{6.4.9}$$

Summing (6.4.8) and (6.4.9) gives an approximation for the entropy production rate. It is this approximate entropy production rate which can efficiently be made positive. Evaluating (6.4.9) at $\phi_{j-\frac{1}{2}} = 1$ gives an approximation for $\overline{(\dot{P}_s^-)_j}$.

$$\Delta x_j \overline{(\dot{P}_s^-)_j} \approx \frac{1}{2} u_j (u_j - u_{j-1})^2 \tag{6.4.10}$$

Using these approximations, it is easy to satisfy the sufficient condition (6.3.5). With very little algebra it can be expressed (for $u_j > 0$) as

$$\phi_{j+\frac{1}{2}} \leq \left(\frac{u_j - u_{j-1}}{u_{j+1} - u_j}\right)^2 \tag{6.4.11}$$

The right-hand side is just the easily computed ${r_j}^2$. An equivalent expression is

$$\phi_{j+\frac{1}{2}} \leq |r_j| \tag{6.4.12}$$

The convention (6.3.3) requires that $\phi_{j+\frac{1}{2}} \leq 1$ when u_j and u_{j+1} are both positive. So in this case, a sufficient constraint for satisfying the approximate entropy inequality is

$$\phi_{j+\frac{1}{2}} \leq \min(1, |r_j|) \tag{6.4.13}$$

It is worth noting that this limiter is identical to the limiter used for the wave equation. In fact the approximation used amounts to assuming a constant wave speed between $u_{j-\frac{1}{2}}$ and $u_{j+\frac{1}{2}}$. This is essentially the same approximation (but not quite the same limiter) that is used in TVD schemes.

Expression (6.4.10) is only valid if u_j and u_{j+1} are positive. If both are negative, the equations are different but the steps are the same. The limiter in this case turns out to be,

$$\phi_{j+\frac{1}{2}} \geq \max\left(1, \left(2 - \left|\frac{1}{r_j}\right|\right)\right) \tag{6.4.14}$$

This expression can also be derived by reversing the coordinate system (so that u_j and u_{j+1} are positive) and then using (6.4.12) directly.

The approximation that $u_{j+1} - u_j$ is small compared to u_j is especially bad at shocks where the former is large and the latter is small. For this reason the first-order scheme is used at shocks.

6.5 Results

If the approximation of a locally constant wave speed is made, this scheme becomes formally TVD. In the numerical experiments which follow, the locally constant wave speed approximation (6.4.13) is used. To check the approximation, $\phi_{j+\frac{1}{2}}$ is also computed without the approximation by using Newton's method. The two values of $\phi_{j+\frac{1}{2}}$ are compared. Where appropriate, the effect of any discrepancy on the solution is noted.

The initial condition chosen here is a two-period sine wave of unit amplitude. This problem contains sonic points and develops shocks. Given that the domain is of length 4π, the shocks will first occur at time $t = 1$, rapidly strengthening until time $t = \frac{\pi}{2}$. At the shock, the exact solution is double valued. After $t = \frac{\pi}{2}$, the shocks diminish steadily in amplitude while the solution maintains its double sawtooth shape. The steady state solution is $u = 0$.

In figure 6.5 the initial condition is plotted. The open diamonds represent the computed solution, while the solid line represents the exact solution. The boundary conditions are Dirichlet, but periodic boundary conditions would give the same results in this case.

Figure 6.6 shows the initial plots of $\phi(x)$. In this case, the solid line represents the values computed with Newton's method (labeled as exact) and the diamonds represent the values computed with (6.4.13) and (6.4.14) (labeled as approximate).

Notice that in regions where $(u^2)_{,x} > 0$, the approximate limiter is just $\phi_{j+\frac{1}{2}} = 1$ which approximates central differences. In compression regions, some dissipation is added, proportionately more as the compression becomes stronger. Here the method approximates the Warming-Beam second-order upwind scheme. In a sense, the scheme simply switches between a second-order upwind scheme and a central difference scheme.

When $|u_j|$ is much larger than $|(u_{j+1} - u_j)|$, (6.4.8) and (6.4.9) are good approximations for $(\dot{P}_s^+)_j$ and $(\dot{P}_s^-)_j$. In these regions, (6.4.13) ($u > 0$) and (6.4.14) ($u < 0$) yield

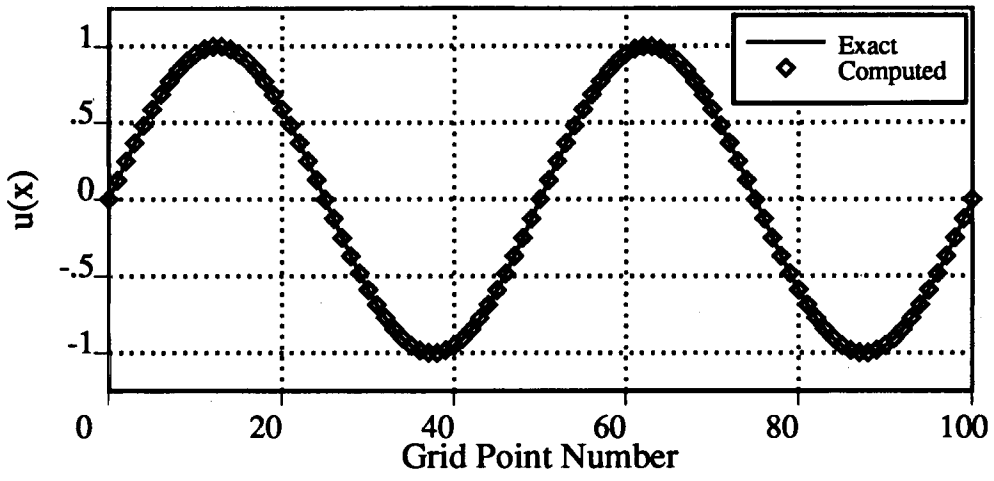

Figure 6.5. – Burgers' equation will sharpen this initial condition into a two-period sawtooth wave as time passes.

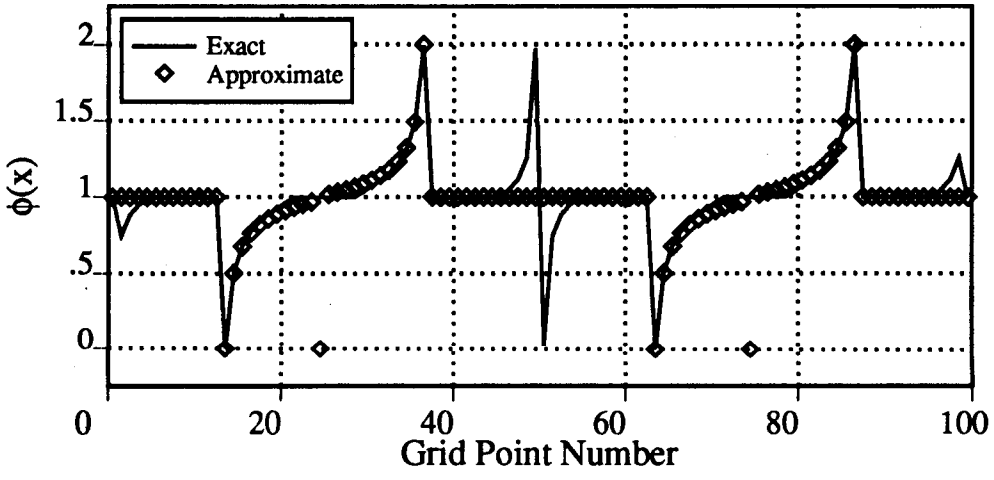

Figure 6.6. – The value $\phi = 1$ corresponds to a symmetric interpolation. Other values indicate various degrees of upwind bias. Two different ways of computing ϕ are compared here for the initial condition.

nearly the same value of $\phi_{j+\frac{1}{2}}$ as the much more costly Newton's method.

This is not true however, when u_j is near zero. Referring to figure 6.6, there is significant disagreement between the two methods in small regions around sonic points and "shocks." These regions become smaller as the mesh is refined.

At $x = \pi$ and $x = 3\pi$ where u changes sign, the exact ϕ is continuous while the

approximate ϕ is obviously discontinuous. At $x = 2\pi$, the sonic point, the reverse occurs. The approximate ϕ is continuous while the exact ϕ is discontinuous.

The difference between the exact and approximate values for ϕ can be explained entirely by the nonlinear variation of wave speed across each cell. Consider a case in which $u(x)$ is linear and positive. The exact entropy condition (6.2.2) reduces to

$$(u_j - u_{j-\frac{1}{2}})^2 \left[u_j - \frac{2}{3}(u_j - u_{j-\frac{1}{2}})\right] - (u_{j+\frac{1}{2}} - u_j)^2 \left[u_j + \frac{2}{3}(u_{j+\frac{1}{2}} - u_j)\right] \geq 0 \qquad (6.5.1)$$

If $(u_{j+\frac{1}{2}} - u_j) = (u_j - u_{j-\frac{1}{2}})$ (i.e., no dissipation), then satisfaction of the inequality depends on the signs of $(u_j - u_{j-\frac{1}{2}})$ and $(u_{j+\frac{1}{2}} - u_j)$. Around the sonic point, both are positive, requiring some adjustment of ϕ to make the magnitudes different. On the other hand, when both are negative, the effect is beneficial, allowing the dissipation to be turned off altogether.

The constant wave speed approximation obscures this effect. With the approximation, the entropy condition reduces to

$$(u_j - u_{j-\frac{1}{2}})^2 - (u_{j+\frac{1}{2}} - u_j)^2 \geq 0 \qquad (6.5.2)$$

Satisfaction of this inequality depends only on the magnitudes of the two terms and not on their signs. Around a sonic point, both terms have approximately the same magnitude so ϕ is adjusted very little, if at all. Around a shock, the first-order scheme is used.

In figure 6.7, the semi-discrete cell entropy production for each cell is given. The symbols show the values for the approximate ϕ while the solid line shows the values obtained from the exact ϕ. These values are independent of the time step. There is very little entropy production initially because there is no shock. What little there is, is numerical and should be eliminated if possible.

The solid line has no negative values, as planned. This shows that a cell entropy inequality can be satisfied in principle. The symbols, on the other hand show negative values near the sonic point, and near the boundaries (which are essentially sonic points). They also show a relatively large amount of entropy production in the cells at $x = \pi$ and $x = 2\pi$ where the first-order method is applied. The accuracy penalty exacted by the approximation is apparent throughout the entire compressive region.

In carrying out the time integration, the scheme

$$u^{n+1} = u^n - \frac{\Delta t}{\Delta x}(f^{n+\frac{1}{2}}_{j+\frac{1}{2}} - f^{n+\frac{1}{2}}_{j-\frac{1}{2}}) \qquad (6.5.2)$$

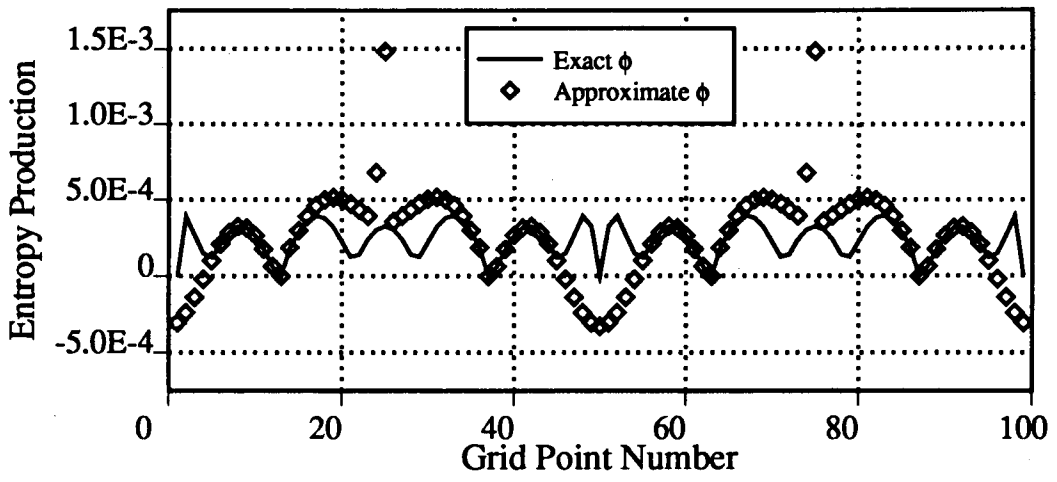

Figure 6.7. – Initial semi-discrete cell entropy production rates. Notice the negative values near the sonic points.

is ideal for scalars. As explained in Section 4.4, the fully discrete entropy production of this scheme is the same as the semi-discrete entropy production, independent of time step. The scheme is very difficult to implement exactly because of linearization errors. As explained in Section 4.5, the following variant is much more practical

$$\begin{aligned} u^{n+\frac{1}{2}} &= u^n - \frac{1}{2}\frac{\Delta t}{\Delta x}(f^*_{j+\frac{1}{2}} - f^*_{j-\frac{1}{2}}) \\ u^{n+1} &= u^n - \frac{\Delta t}{\Delta x}(f^{n+\frac{1}{2}}_{j+\frac{1}{2}} - f^{n+\frac{1}{2}}_{j-\frac{1}{2}}) \end{aligned} \tag{6.5.3}$$

The fluxes $f^*_{j+\frac{1}{2}}$ and $f^*_{j-\frac{1}{2}}$ are computed with some suitable linearization so as to approximate $f^{n+\frac{1}{2}}_{j+\frac{1}{2}}$ and $f^{n+\frac{1}{2}}_{j-\frac{1}{2}}$. Computing derivatives such as $\frac{\partial f^*_{j+\frac{1}{2}}}{\partial q_j}$ requires derivatives such as $\frac{\partial \phi_{j+\frac{1}{2}}}{\partial q_j}$. This is where equations (6.4.12) and (6.4.14) exhibit a tremendous computational advantage. The approximate definition of $\phi_{j+\frac{1}{2}}$ is piecewise linear and derivatives can be obtained analytically. By contrast, if $\phi_{j+\frac{1}{2}}$ is computed with Newton's method, numerical differentiation is needed. For this reason, the approximate ϕ values are used to advance in time. The maximum CFL number used will be 0.5.

In figure 6.8, the solution is plotted at time $t \approx 1.005$. The sharpening of the solution into a shock is quite apparent. There are no oscillations anywhere and the computed solution is tracking the exact solution quite well. There is a slight loss in amplitude at the peak values, perhaps due to the excess dissipation in the compressive regions.

In figure 6.9, the values of ϕ are shown at the time of figure 6.8. Clearly the compressive region has shrunk and become more intense, while the expansion region now occupies

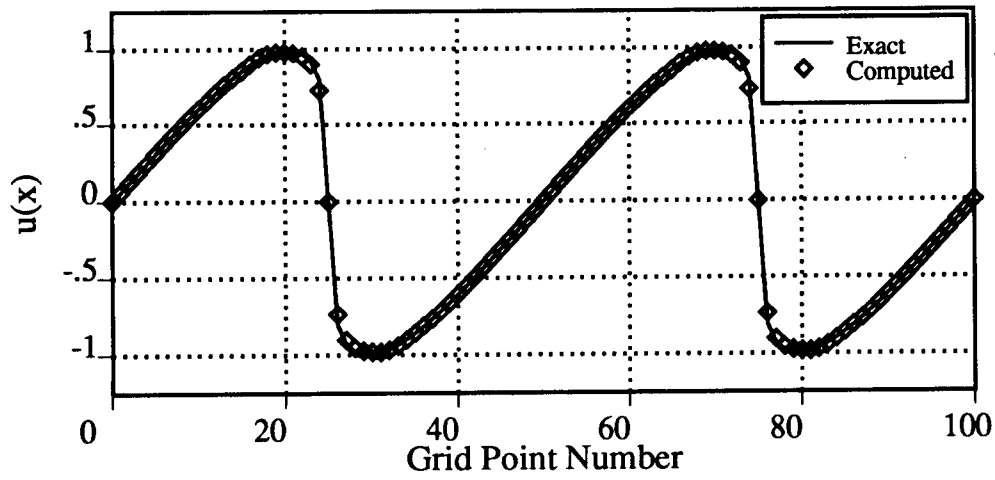

Figure 6.8. – Onset of the shock.

most of the domain. Notice how the solid line has begun to respond (symmetrically) to the shock. Furthermore, the approximate ϕ is becoming more symmetric as it responds to what it perceives as a very strong compression next to a shock. In fact, there is no way to distinguish between a strong compression and a shock on a finite mesh.

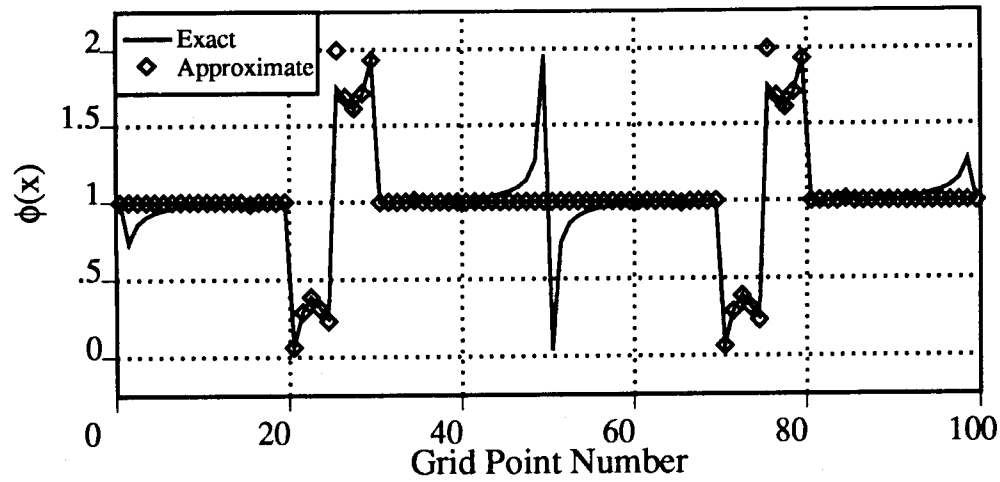

Figure 6.9. – Dissipation at shock onset.

In figure 6.10, the semi-discrete entropy production is shown. Notice the larger values of entropy production that have begun to appear at the shock (the scale has changed by three orders of magnitude). Also notice the way in which the dissipation has followed the

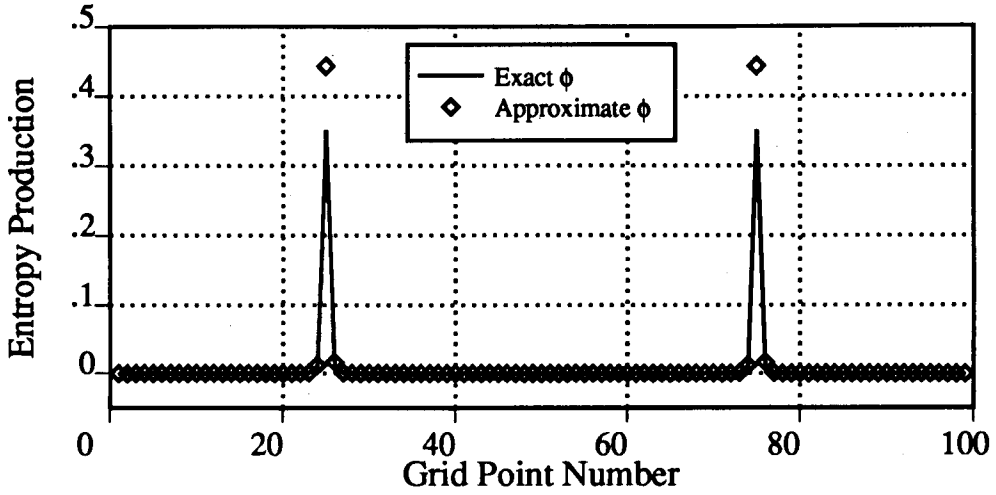

Figure 6.10. – With the appearance of the shock comes significant entropy production within the shock cells. For example, compare scales with figure 6.7.

solution so that essentially all of the dissipation is at the shock as required by the physics. Finally, the entropy production rate is higher than it should be.

At $t = \frac{\pi}{2}$ the peak of the sine wave has traveled exactly 12.5 grid points and the peak shock strength occurs. The computed solution is shown in figure 6.11. The computed solution is very smooth and tracks the exact solution except at the shocks. There is again a slight loss of amplitude.

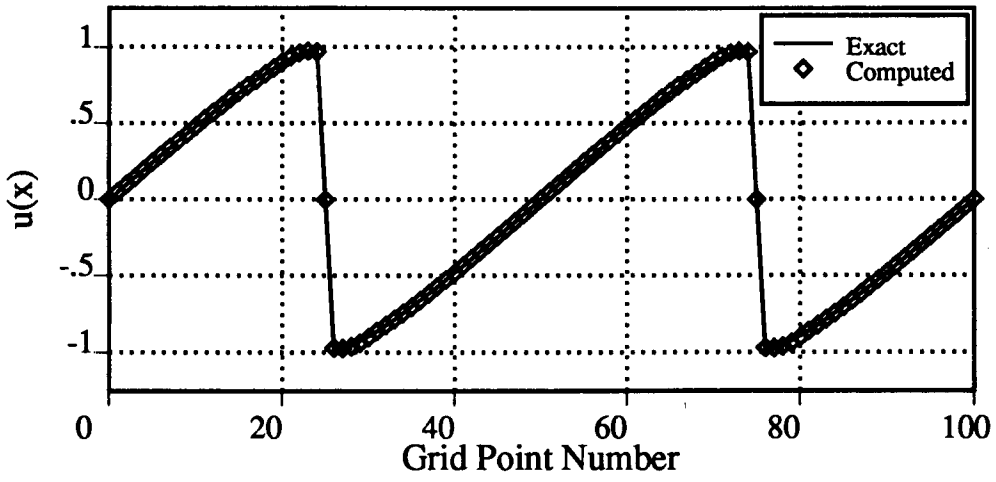

Figure 6.11. – This is as strong as the shock gets. After this it slowly decays to zero.

In figure 6.12, the compressive regions have shrunk to a two-cell interval in the neighborhood of the shock. The remainder of the domain contains expansion regions. Notice that $\phi_{j+\frac{1}{2}} = 1$ except at the shock. This corresponds to the analytic solution in that all of the dissipation occurs at the shock.

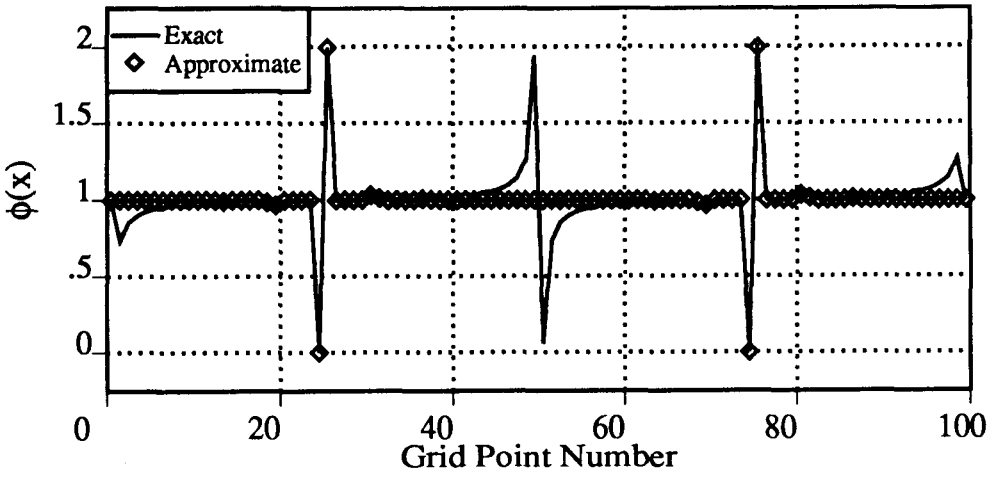

Figure 6.12. – Essentially all of the dissipation introduced by the numerical method is at the shock. This corresponds to the physical situation.

The peak shock strength can be computed analytically. As shown in figure 6.13, the peak entropy production rate for the cell containing the shock is exactly $\frac{4}{3}$. The corresponding numerical shock strength (represented by the diamonds) is about 1.21, a significant error. This is due to numerical dissipation inherent in the method. Entropy production rates away from the shock are less than 10^{-6} in magnitude, though, as previously noted, some of them are negative.

Interestingly, the computed shock strength does not decay quite as fast as the exact shock strength. Indeed, the numerical shocks eventually become stronger than the exact ones. Perhaps this is due to the entropy violations observed early in the calculation. These influences would not be felt at the shock until about $t = \pi$ which is about what is observed. Note that the entropy production errors at sonic points do not jeopardize the stability of the computation. Since the total entropy produced in the domain is positive, the stability proof of Chapter 2 applies and the solution remains bounded.

It is now reasonable to review the costs and benefits of the locally constant wave speed approximation. This approximation results in a scheme that is more dissipative than it needs to be during transient compressions, though the effect is not severe. It also results in minor violations of the second law in the neighborhood of sonic points. These

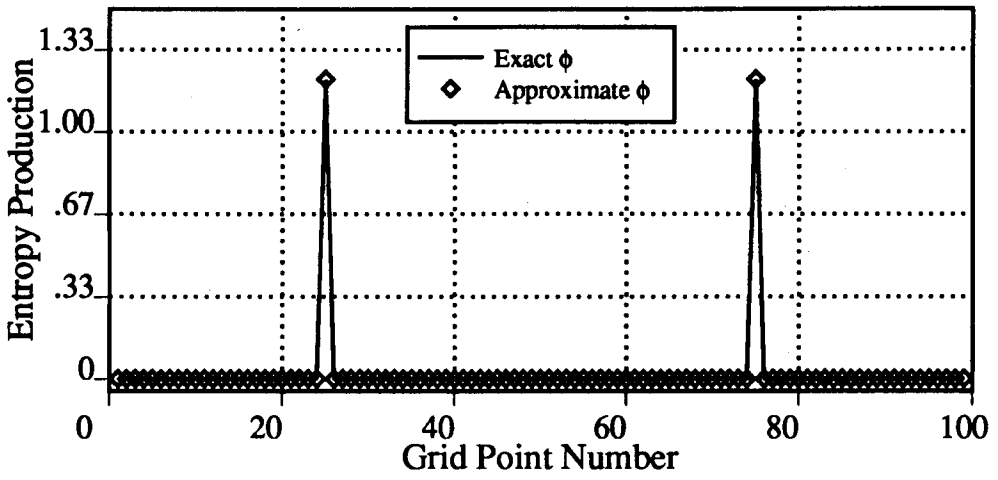

Figure 6.13. – The analytic solution reaches a peak entropy production rate of 1.33 . The shortfall indicates accumulated error.

errors can contaminate the solution elsewhere but the effect is small. On the other hand, the computational convenience of such an approximation is very significant. The expense of using Newton's method to compute $\phi_{j+\frac{1}{2}}$ is quadrupled by the requirement of using numerical linearizations. In sum, the practical advantages of the approximation more than make up for its theoretical shortcomings.

The time advance scheme used provides at least an *a posteriori* evaluation of the fully discrete entropy production. In this way the importance of the linearization errors can be glimpsed. In figures 6.14 and 6.15, the solid line represents the fully discrete entropy production, that is the total entropy produced in a given cell over the time step divided by the size of the time step. The symbols represent the semi-discrete entropy production halfway through the first step. figure 6.14 corresponds to a CFL number of 2. Notice that the fully discrete inequality corresponds very well with the semi-discrete scheme. This shows that the linearization errors are not too important at this time increment. At smaller time steps the situation improves further. At a CFL number of $\frac{1}{2}$ there is no discernable difference.

On the other hand, figure 6.15 corresponds to a CFL number of 5 which is distinctly unstable. The fully discrete entropy production rates are significantly different than the semi-discrete values. Some violations of the second law are apparent even in the first step. This indicates large errors due to linearization in time. If (6.5.2) could somehow be implemented exactly, instead of in a linearized sense, the scheme would be unconditionally stable.

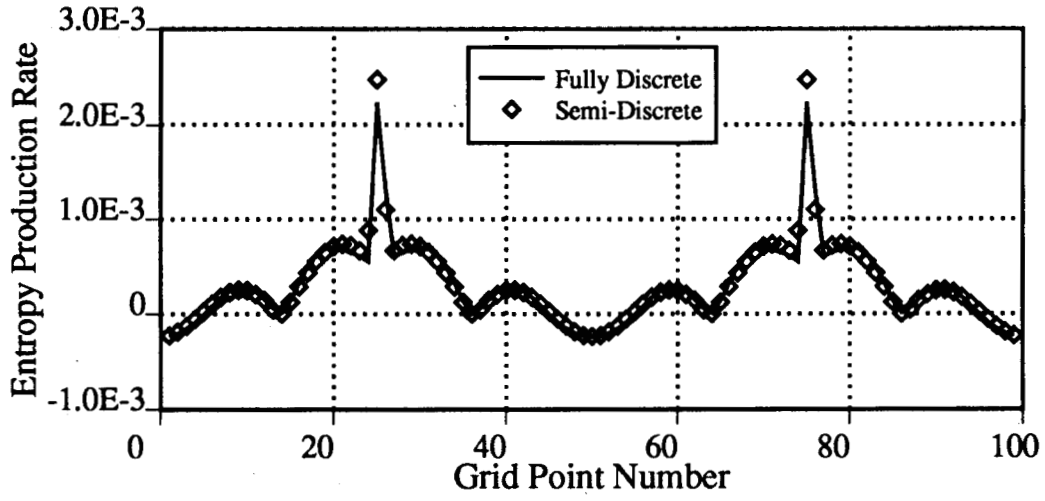

Figure 6.14. – At CFL = 2, the semi-discrete entropy production rates are in good agreement with the fully discrete rates. This indicates that the time linearization errors are small.

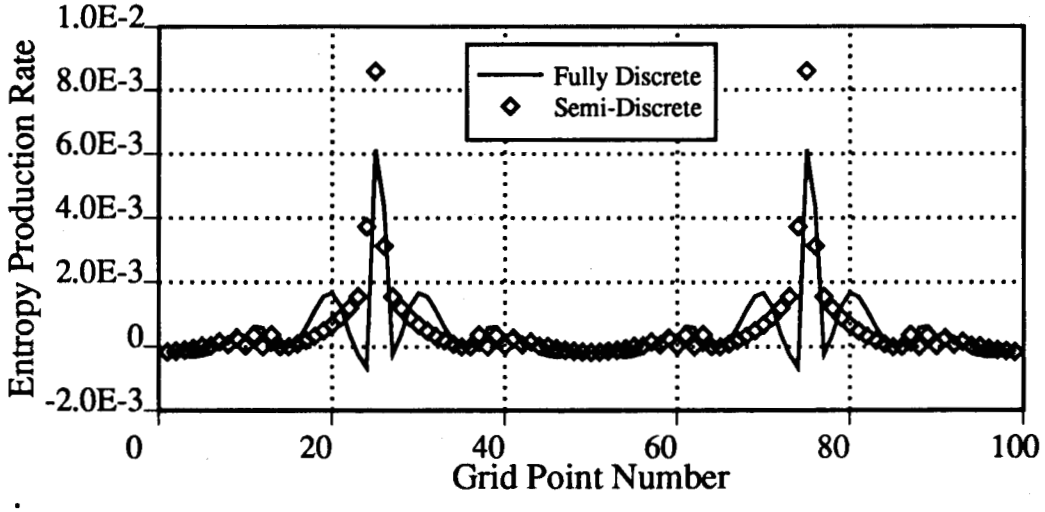

Figure 6.15. – At CFL = 5, the semi-discrete entropy production rates are no longer in agreement with the fully discrete rates. Negative entropy production rates are evident in several cells. This indicates that the time linearization errors are significant. To reduce them requires a smaller time step or iteration within a time step.

Burgers' equation provides a good forum for testing approximations and implementing schemes. In this chapter, the problems of shocks and sonic points have been explored and the approximation of constant wave speed has been investigated. The effect of a finite time

step in the presence of linearization errors has also been investigated. Most importantly, it has been shown that a semi-discrete cell entropy inequality can be systematically satisfied, at least for the scalar case.

In the next chapter these ideas will be extended to systems of equations. This will introduce several additional complications, among them: multiple wave speeds, definition of ϕ and linearization difficulties. The fundamental principles remain unchanged. Satisfaction of a fully discrete cell entropy inequality (plus consistency and conservation) guarantees a stable and accurate solution. No other paradigm can promise this.

Chapter 7. ONE-DIMENSIONAL GASDYNAMICS

7.1 The One-Dimensional Euler Equations

This chapter is concerned with the numerical solution of the one-dimensional Euler equations. These were introduced in Chapter 2. For convenience, selected portions will be repeated here.

In one dimension the Euler equations can be written

$$q_{,t} + f_{,x} = 0 \tag{7.1.1}$$

where

$$q = \begin{bmatrix} \rho \\ \rho u \\ e \end{bmatrix} \quad \text{and} \quad f = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ (e+p)u \end{bmatrix} \tag{7.1.2}$$

The quantities ρ and u represent the local mass density and the local velocity, respectively. The quantity e represents the total energy density. An expression for the pressure p is

$$p = (\gamma - 1)(e - \frac{1}{2}\rho u^2) \tag{7.1.3}$$

A suitable entropy pair for inviscid, one-dimensional gasdynamics is

$$S = \rho s \tag{7.1.4}$$

$$F = \rho u s \tag{7.1.5}$$

where

$$s = \log(\frac{p}{\rho^\gamma}) \tag{7.1.6}$$

As shown in Section 2.3, this pair agrees with the physical viewpoint and satisfies the convexity and compatability conditions. When the problem is made more complicated by the addition of viscosity, heat conduction, and three-dimensional domains, the entropy does not change at all and the entropy flux changes only slightly. Although this chapter (and indeed, this entire work) does not proceed beyond the simple case of one-dimensional inviscid flow, there is little conceptual difficulty in doing so. This is discussed further in Chapter 10.

7.2 Entropy Variables

The main difference between equations (7.1.1) and equation (6.1.1) is that q is a vector and u is a scalar. This introduces a number of ambiguities which need to be discussed. The first issue to be decided is the choice of independent variables. A common choice is the conservative variables.

$$q \equiv \begin{bmatrix} \rho \\ \rho u \\ e \end{bmatrix} \tag{7.2.1}$$

These are called the conservative variables because domain integrals of them are conserved physically (except for boundary conditions). They facilitate construction of schemes which conserve mass momentum and energy numerically.

Other variables are equally valid. For example, primitive variables are often used because of the ease with which they can be measured experimentally. These are

$$q' \equiv \begin{bmatrix} \rho \\ u \\ p \end{bmatrix} \tag{7.2.2}$$

There is a one-to-one correspondence between the conservative variables and the primitive variables. That is, each set can be computed from the other. If computational expense is not an issue, any scheme which can be implemented in one set of variables can be implemented in the other, using two conversions per step.

The issue of conservation is independent of the choice of variables. A scheme conserves mass, momentum and energy if the computed fluxes are the conservative ones

$$f \equiv \begin{bmatrix} \rho u \\ \rho u^2 + p \\ (e+p)u \end{bmatrix} \tag{7.2.3}$$

and if these are evaluated on the control volume boundaries. These fluxes can be evaluated with equal ease using either the conservative or the primitive variables.

From this point of view, any set of variables will do, provided there is a one-to-one correspondence with the variables of (7.2.1). This is sufficient to ensure that the conservative fluxes can be computed and conservative schemes constructed. The issues of consistency and satisfaction of the second law remain.

In Chapter 2 the definition of consistency was given. If information at several grid points is used to compute a flux, and if the values at those grid points are all the same, then the computed flux must match the flux at the grid points. This is a fairly easy criterion to meet. In this work, it is met by requiring that $q_{j+\frac{1}{2}}$ be between q_j and q_{j+1}.

In the case of scalars, it is clear what "between" means. The value $u_{j+\frac{1}{2}}$ is said to be between u_j and u_{j+1} if

$$(u_{j+1} - u_{j+\frac{1}{2}})(u_{j+\frac{1}{2}} - u_j) \geq 0 \tag{7.2.4}$$

In the case of systems, the definition is less obvious. In this chapter, $q_{j+\frac{1}{2}}$ is said to be between q_j and q_{j+1} if

$$\left(\frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}}\right) \cdot \left(\frac{\partial(\dot{P}_s^-)_{j+1}}{\partial q_{j+\frac{1}{2}}}\right) \geq 0 \tag{7.2.5}$$

As shown in Section 3.4, (7.2.5) reduces to (7.2.4) in the case of scalars.

The expressions for $(\dot{P}_s^+)_j$ and $(\dot{P}_s^-)_{j+1}$ are (as derived in Section 3.4)

$$\Delta x_j(\dot{P}_s^+)_j = -(S_{,q})_j(f_{j+\frac{1}{2}} - f_j) + (F_{j+\frac{1}{2}} - F_j) \tag{7.2.6}$$

$$\Delta x_{j+1}(\dot{P}_s^-)_{j+1} = -(S_{,q})_{j+1}(f_{j+1} - f_{j+\frac{1}{2}}) + (F_{j+1} - F_{j+\frac{1}{2}}) \tag{7.2.7}$$

and the gradients used in (7.2.5) are

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}} = \left[(S_{,q})_{j+\frac{1}{2}} - (S_{,q})_j\right](f_{,q})_{j+\frac{1}{2}} \tag{7.2.8}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial q_{j+\frac{1}{2}}} = \left[(S_{,q})_{j+1} - (S_{,q})_{j+\frac{1}{2}}\right](f_{,q})_{j+\frac{1}{2}} \tag{7.2.9}$$

Suppose that $q_{j+1} = q_j$. By inspection, the two gradients have the same magnitude but opposite sign. Condition (7.2.5) only holds if both gradients vanish, that is if $(S_{,q})_j = (S_{,q})_{j+\frac{1}{2}} = (S_{,q})_{j+1}$.

Remarkably, this is sufficient to imply that $q_j = q_{j+\frac{1}{2}} = q_{j+1}$. Let $S_{,q}$ define a new set of dependent variables, v, that is, from (2.3.13)

$$v \equiv {S_{,q}}^T = \left[\, s - (\gamma+1) + \tfrac{(\gamma-1)e}{p}, \quad -\tfrac{(\gamma-1)\rho u}{p}, \quad \tfrac{(\gamma-1)\rho}{p} \quad \right]^T \tag{7.2.10}$$

Since $v_{,q} = S_{,qq}$, and by convexity $S_{,qq} < 0$, it is clear that $v_{,q} \neq 0$. In general, this is sufficient to ensure a one-to-one mapping between q and v. In particular, for the one-dimensional Euler equations, v can be computed from q using (7.2.10) and q can be computed from v by computing u,s,ρ,p, and e in the following sequence:

$$u = -\frac{v_2}{v_3} \tag{7.2.11}$$

$$s = v_1 + \gamma + \frac{1}{2}uv_2 \tag{7.2.12}$$

$$\rho = \left[\frac{(\gamma-1)}{v_3 \exp(s)}\right]^{\frac{1}{(\gamma-1)}} \tag{7.2.13}$$

$$p = \frac{(\gamma-1)\rho}{v_3} \tag{7.2.14}$$

$$e = \frac{p}{(\gamma-1)} + \frac{1}{2}\rho u^2 \tag{7.2.15}$$

where the so called "entropy variables," v_1, v_2, and v_3 are the elements of v as shown in (7.2.10). This transformation is due to Mock (ref. 26).

To summarize, if $q_j = q_{j+1}$, and condition (7.2.5) is met, then $q_j = q_{j+\frac{1}{2}} = q_{j+1}$. It follows that $f_j = f_{j+\frac{1}{2}} = f_{j+1}$. Thus (7.2.5) is sufficient to assure consistency in the sense of Lax. Careful examination of (7.2.8) and (7.2.9) (the factors of (7.2.5)) shows that they are easily written in terms of entropy variables (see Section 7.4). Thus, it is useful to think in terms of the entropy variables v when constructing consistent schemes.

7.3 A Remarkable Identity

In constructing schemes that satisfy an entropy inequality, the basic approach for systems is to diagonalize the equations locally. This can be done without approximation for the one-dimensional Euler equations, mostly because of the existence of a remarkable identity.

Recall from Section 2.3 that the convexity condition was demonstrated for these equations using the factorization

$$S_{,qq} = -(Y^{-1})^T(Y^{-1}) \tag{7.3.1}$$

It is well known that the flux Jacobian also has a factorization

$$f_{,q} = Y\Lambda Y^{-1} \tag{7.3.2}$$

where

$$\Lambda = \begin{bmatrix} u & & \\ & u-c & \\ & & u+c \end{bmatrix} \tag{7.3.3}$$

contains the eigenvalues of $f_{,q}$ and the columns of Y contains the eigenvectors of $f_{,q}$. It is well known that each eigenvector is unique only up to a multiplicative constant. It is possible to choose these constants in such a way that (7.3.1) holds, using the same Y as (7.3.2).

The matrices Y and Y^{-1} are repeated here with the correct scaling. The speed of sound, $c = \sqrt{\frac{\gamma p}{\rho}}$, has been used to simplify the expressions.

$$Y = \begin{bmatrix} \beta_1 & \beta_2 & \beta_2 \\ \beta_1 u & \beta_2(u-c) & \beta_2(u+c) \\ \frac{\beta_1 u^2}{2} & \beta_2(\frac{u^2}{2} - uc + \frac{c^2}{(\gamma-1)}) & \beta_2(\frac{u^2}{2} + uc + \frac{c^2}{(\gamma-1)}) \end{bmatrix} \tag{7.3.4}$$

$$Y^{-1} = \left(\frac{1}{2c^2}\right) \begin{bmatrix} \frac{2c^2-(\gamma-1)u^2}{\beta_1} & \frac{2(\gamma-1)u}{\beta_1} & \frac{-2(\gamma-1)}{\beta_1} \\ \frac{(\gamma-1)u^2}{2\beta_2} + \frac{uc}{\beta_2} & \frac{-(\gamma-1)u-c}{\beta_2} & \frac{(\gamma-1)}{\beta_2} \\ \frac{(\gamma-1)u^2}{2\beta_2} - \frac{uc}{\beta_2} & \frac{-(\gamma-1)u+c}{\beta_2} & \frac{(\gamma-1)}{\beta_2} \end{bmatrix} \tag{7.3.5}$$

$$\beta_1 = \sqrt{\frac{\rho}{\gamma}}, \quad \beta_2 = \sqrt{\frac{\rho}{2\gamma(\gamma-1)}}$$

The fact that β_2 is used twice is not surprising. The equations are expected to be invariant to a reversal of the coordinate system, so the $u + c$ and $u - c$ waves are scaled the same way.

The surprising thing is that $(S_{,qq})^{-1}$ can be decomposed in terms of the eigenvectors of $f_{,q}$. This is a new result. It has been verified that this is also true for the shallow water equations in one dimension (a two-by-two system)[1], and of course, for scalar equations. So far, no theoretical explanation is offered. The identity itself was found and verified using MACSYMA.

There are some interesting corollaries. It will be necessary in the next section to deal with the flux Jacobian with respect to v, the entropy variables. This is related to the usual flux Jacobian and its eigensystem as follows:

$$f_{,q} = f_{,v} v_q \qquad \text{chain rule} \qquad (7.3.6)$$

$$= f_{,v} S_{,qq} \qquad \text{definition of } v \qquad (7.3.7)$$

$$Y \Lambda Y^{-1} = -f_{,v} (Y^{-1})^T (Y^{-1}) \qquad (7.3.1) \text{ and } (7.3.2) \qquad (7.3.8)$$

$$f_{,v} = -Y \Lambda Y^T \qquad \text{from } (7.3.8) \qquad (7.3.9)$$

Equation (7.3.9) is consistent with the well known symmetry of $f_{,v}$.

The term $S_{,qq} f_{,q}$ also appears from time to time. Using (7.3.1) and (7.3.2) immediately gives

$$S_{,qq} f_{,q} = (Y^{-1})^T \Lambda (Y^{-1}) \qquad (7.3.10)$$

which is also a symmetric form.

In summary, identity (7.3.1) provides a new tool with which to study hyperbolic systems of equations in the context of the second law. In the next section it allows the equations to be diagonalized without approximation.

7.4 A Second-Order Scheme

For this problem, construction of a first-order accurate scheme which satisfies a semi-discrete cell entropy inequality is nearly as difficult as constructing a second-order accurate scheme with this property. For this reason, the first-order scheme is left to the interested reader.

In this section, a second-order accurate scheme is described, which also satisfies a semi-discrete cell entropy inequality. The method presented here is analogous to the one for Burgers' equation, but with additional complications. The primary emphasis in Chapter 6 was on suitable approximations to make the computation more practical. In this chapter, the emphasis is on rigorously satisfying the semi-discrete entropy inequality, without regard for the practical aspects. The use of implicit Euler time advance allows the semi-discrete entropy inequality to be extended to a fully discrete one as explained in Section 4.3.

[1] P.L. Roe, personal communication

Recall that the consistency condition is

$$\left(\frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}}\right) \cdot \left(\frac{\partial(\dot{P}_s^-)_{j+1}}{\partial q_{j+\frac{1}{2}}}\right) \geq 0 \tag{7.4.1}$$

The gradients, previously given in terms of q ((7.2.8) and (7.2.9)), can be restated as

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}} = \left[v_{j+\frac{1}{2}} - v_j\right]^T (f_{,q})_{j+\frac{1}{2}} \tag{7.4.2}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial q_{j+\frac{1}{2}}} = \left[v_{j+1} - v_{j+\frac{1}{2}}\right]^T (f_{,q})_{j+\frac{1}{2}} \tag{7.4.3}$$

Consistency only requires that (7.4.1) hold when $q_j = q_{j+1}$. However, other aspects of the scheme require (7.4.1) to hold more generally. By inspection of (7.4.2) and (7.4.3), a special case occurs when $v_{j+\frac{1}{2}}$ is chosen to be the arithmetic mean, $\bar{v} \equiv \frac{1}{2}(v_j + v_{j+1})$. This choice has the property that

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial q_{j+\frac{1}{2}}} = \Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial q_{j+\frac{1}{2}}} \tag{7.4.4}$$

which certainly satisfies (7.4.1). Of course $\bar{q} \equiv q(\bar{v})$ also does.

Given that $\bar{q}$ is defined in terms of v, it seems reasonable to define the gradients in terms of v as well. The algebra for this requires the compatability condition in terms of entropy variables. This is derived as follows:

$$S_{,q} f_{,q} = f_{,q} \qquad \text{compatibility condition} \tag{7.4.5}$$

$$S_{,q} f_{,q} q_{,v} = f_{,q} q_{,v} \tag{7.4.6}$$

$$S_{,q} f_{,v} = F_{,v} \qquad \text{chain rule} \tag{7.4.7}$$

Recall that $(\dot{P}_s^+)_j$ and $(\dot{P}_s^-)_{j+1}$ are given by (7.2.6) and (7.2.7). Differentiating these with respect to $v_{j+\frac{1}{2}}$ gives

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial v_{j+\frac{1}{2}}} = -(S_{,q})_j (f_{,v})_{j+\frac{1}{2}} + (F_{,v})_{j+\frac{1}{2}} \tag{7.4.8}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial v_{j+\frac{1}{2}}} = (S_{,q})_{j+1} (f_{,v})_{j+\frac{1}{2}} - (F_{,v})_{j+\frac{1}{2}} \tag{7.4.9}$$

Now applying the compatability condition in entropy variables (7.4.7) gives

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial v_{j+\frac{1}{2}}} = -(S_{,q})_j (f_{,v})_{j+\frac{1}{2}} + (S_{,q})_{j+\frac{1}{2}} (f_{,v})_{j+\frac{1}{2}} \tag{7.4.10}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial v_{j+\frac{1}{2}}} = (S_{,q})_{j+1} (f_{,v})_{j+\frac{1}{2}} - (S_{,q})_{j+\frac{1}{2}} (f_{,v})_{j+\frac{1}{2}} \tag{7.4.11}$$

Collecting terms gives

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial v_{j+\frac{1}{2}}} = \left[(S_{,q})_{j+\frac{1}{2}} - (S_{,q})_j\right](f_{,v})_{j+\frac{1}{2}} \tag{7.4.12}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial v_{j+\frac{1}{2}}} = \left[(S_{,q})_{j+1} - (S_{,q})_{j+\frac{1}{2}}\right](f_{,v})_{j+\frac{1}{2}} \tag{7.4.13}$$

Finally, using the definition of v (7.2.10), gives

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial v_{j+\frac{1}{2}}} = (v_{j+\frac{1}{2}} - v_j)^T (f_{,v})_{j+\frac{1}{2}} \tag{7.4.14}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial v_{j+\frac{1}{2}}} = (v_{j+1} - v_{j+\frac{1}{2}})^T (f_{,v})_{j+\frac{1}{2}} \tag{7.4.15}$$

Again, if $v_{j+\frac{1}{2}} = \bar{v}$,

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial v_{j+\frac{1}{2}}} = \Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial v_{j+\frac{1}{2}}} \tag{7.4.16}$$

which is analogous to (7.4.4). It is, however, important to explore the behavior of these gradients away from $v_{j+\frac{1}{2}} = \bar{v}$. Identity (7.3.9) is very helpful in this regard.

$$f_{,v} = -Y \Lambda Y^T \tag{7.4.17}$$

Recall that Λ and Y are the eigenvalues and eigenvectors of $f_{,q}$ (not $f_{,v}$!) Using this identity, equations (7.4.14) and (7.4.15) may be rewritten

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial v_{j+\frac{1}{2}}} = -(v_{j+\frac{1}{2}} - v_j)^T Y \Lambda Y^T \tag{7.4.18}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial v_{j+\frac{1}{2}}} = -(v_{j+1} - v_{j+\frac{1}{2}})^T Y \Lambda Y^T \tag{7.4.19}$$

where both Y and Λ are evaluated at $q(v_{j+\frac{1}{2}})$. Bringing Y inside the parenthesis gives

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial v_{j+\frac{1}{2}}} = -(Y^T v_{j+\frac{1}{2}} - Y^T v_j)^T \Lambda Y^T \tag{7.4.20}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial v_{j+\frac{1}{2}}} = -(Y^T v_{j+1} - Y^T v_{j+\frac{1}{2}})^T \Lambda Y^T \tag{7.4.21}$$

It is convenient to define characteristic variables.

$$\tilde{v} \equiv Y^T v^T \tag{7.4.22}$$

where Y^T is always evaluated at the still unknown state, $q(v_{j+\frac{1}{2}})$. For example, (7.4.22) implies $\tilde{v}_j = Y^T v j^T$ even though Y^T is evaluated at $v_{j+\frac{1}{2}}$. With this definition (7.4.20) and (7.4.21) may be rewritten as

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial v_{j+\frac{1}{2}}} = -(\tilde{v}_{j+\frac{1}{2}} - \tilde{v}_j)^T \Lambda Y^T \tag{7.4.23}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial v_{j+\frac{1}{2}}} = -(\tilde{v}_{j+1} - \tilde{v}_{j+\frac{1}{2}})^T \Lambda Y^T \tag{7.4.24}$$

Now, using the chain rule, the left-hand side can be rewritten in terms of gradients with respect to the characteristic variables.

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial \tilde{v}_{j+\frac{1}{2}}} \frac{\partial \tilde{v}_{j+\frac{1}{2}}}{\partial v_{j+\frac{1}{2}}} = -(\tilde{v}_{j+\frac{1}{2}} - \tilde{v}_j)^T \Lambda Y^T \tag{7.4.25}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial \tilde{v}_{j+\frac{1}{2}}} \frac{\partial \tilde{v}_{j+\frac{1}{2}}}{\partial v_{j+\frac{1}{2}}} = -(\tilde{v}_{j+1} - \tilde{v}_{j+\frac{1}{2}})^T \Lambda Y^T \tag{7.4.26}$$

Definition (7.4.22) can be used to carry out the differentiation and show that

$$\frac{\partial \tilde{v}_{j+\frac{1}{2}}}{\partial v_{j+\frac{1}{2}}} = Y^T \tag{7.4.27}$$

This leads to the expressions

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial \tilde{v}_{j+\frac{1}{2}}} Y^T = -(\tilde{v}_{j+\frac{1}{2}} - \tilde{v}_j)^T \Lambda Y^T \tag{7.4.28}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial \tilde{v}_{j+\frac{1}{2}}} Y^T = -(\tilde{v}_{j+1} - \tilde{v}_{j+\frac{1}{2}})^T \Lambda Y^T \tag{7.4.29}$$

Both sides are right multiplied by $(Y^T)^{-1}$ leaving the expression

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial \tilde{v}_{j+\frac{1}{2}}} = -(\tilde{v}_{j+\frac{1}{2}} - \tilde{v}_j)^T \Lambda \tag{7.4.30}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial \tilde{v}_{j+\frac{1}{2}}} = -(\tilde{v}_{j+1} - \tilde{v}_{j+\frac{1}{2}})^T \Lambda \tag{7.4.31}$$

Notice that the gradients are now defined with respect to $\tilde{v}$ instead of v. The unknown quantity $\tilde{v}_{j+\frac{1}{2}}$ is now interpolated from the endpoints using the formula (analogous to (6.2.5))

$$\tilde{v}_{j+\frac{1}{2}} = \tilde{v}_j + \frac{1}{2} \phi_{j+\frac{1}{2}}{}^D (\tilde{v}_{j+1} - \tilde{v}_j) \tag{7.4.32}$$

where the notation ${\phi_{j+\frac{1}{2}}}^D$ indicates a diagonal matrix whose entries are the entries of the three component vector $\phi_{j+\frac{1}{2}}$. These components, ϕ_1 ,ϕ_2 , and ϕ_3 vary in value between zero and two. Consistency immediately follows if $\tilde{v}_{j+\frac{1}{2}}$ is chosen using (7.4.32). Using (7.4.32), equations (7.4.30) and (7.4.31) can be rewritten as

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial \tilde{v}_{j+\frac{1}{2}}} = -\frac{1}{2}(\tilde{v}_{j+1} - \tilde{v}_j)^T {\phi_{j+\frac{1}{2}}}^D \Lambda \tag{7.4.33}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial \tilde{v}_{j+\frac{1}{2}}} = -\frac{1}{2}(\tilde{v}_{j+1} - \tilde{v}_j)^T (2I - {\phi_{j+\frac{1}{2}}}^D)\Lambda \tag{7.4.34}$$

When ${\phi_{j+\frac{1}{2}}}^D = I$, the two gradients are equal. As in the scalar version of this scheme, $\phi_{j+\frac{1}{2}}$ may be perturbed from this value to increase the entropy production rate, $(\dot{P}_s)_j$ or $(\dot{P}_s)_{j+1}$. If a given change in an entropy production rate is required, the smallest possible change (in the sense of Euclidean length) in $\phi_{j+\frac{1}{2}}$ is desired. Such changes occur along the gradients of entropy production with respect to $\phi_{j+\frac{1}{2}}$. These can be computed as follows: first note that (7.4.32) may be written

$$\tilde{v}_{j+\frac{1}{2}} = \tilde{v}_j + \frac{1}{2}(\tilde{v}_{j+1} - \tilde{v}_j)^D \phi_{j+\frac{1}{2}} \tag{7.4.35}$$

This implies that

$$\frac{\partial \tilde{v}_{j+\frac{1}{2}}}{\partial \phi_{j+\frac{1}{2}}} = \frac{1}{2}(\tilde{v}_{j+1} - \tilde{v}_j)^D \tag{7.4.36}$$

The chain rule implies that

$$\frac{\partial \dot{P}_s}{\partial \phi_{j+\frac{1}{2}}} = \frac{\partial \dot{P}_s}{\partial \tilde{v}_{j+\frac{1}{2}}} \cdot \frac{\partial \tilde{v}_{j+\frac{1}{2}}}{\partial \phi_{j+\frac{1}{2}}} \tag{7.4.37}$$

Using (7.4.33) or (7.4.34) for the first factor and (7.4.36) for the second yields

$$\Delta x_j \frac{\partial(\dot{P}_s^+)_j}{\partial \phi_{j+\frac{1}{2}}} = -\frac{1}{4}(\tilde{v}_{j+1} - \tilde{v}_j)^T {\phi_{j+\frac{1}{2}}}^D \Lambda (\tilde{v}_{j+1} - \tilde{v}_j)^D \tag{7.4.38}$$

$$\Delta x_{j+1} \frac{\partial(\dot{P}_s^-)_{j+1}}{\partial \phi_{j+\frac{1}{2}}} = -\frac{1}{4}(\tilde{v}_{j+1} - \tilde{v}_j)^T (2I - {\phi_{j+\frac{1}{2}}}^D)\Lambda (\tilde{v}_{j+1} - \tilde{v}_j)^D \tag{7.4.39}$$

The right side of (7.4.38) is a three component row vector, the first component of which is $-\frac{1}{4}(\tilde{v}_{j+1} - \tilde{v}_j)_1^2 \phi_1 \lambda_1$. Consistency requires $0 \leq \phi_1 \leq 2$. The only factor which can be negative is λ_1, one of the local wave speeds and the first component of Λ. This implies that ϕ_1 may have to be reduced in magnitude if $\lambda_1 > 0$, or increased if $\lambda_1 < 0$. Thus the concept of upwind differencing is recovered from the second law, for systems of equations. No approximations were needed for this result. Note that this result does not constitute a scheme; although the gradients are exact, they only give a direction and an approximate magnitude for perturbations to a given vector $\phi_{j+\frac{1}{2}}$.

The matrix ${\phi_{j+\frac{1}{2}}}^D$ must have the same sign as $2I - {\phi_{j+\frac{1}{2}}}^D$, term by term, since the entries are constrained to lie between 0 and 2. Since Λ is evaluated at $q_{j+\frac{1}{2}} = q(v_{j+\frac{1}{2}})$, both gradients must both have the same sign, term by term.

All this information on how the entropy production rate depends on $\phi_{j+\frac{1}{2}}$, is needed to construct a second-order accurate, consistent scheme which satisfies a cell entropy inequality. A semi-discrete expression for the entropy production rate is

$$\Delta x_j(\dot{P}_s)_j = \Delta x_j(\dot{P}_s^+)_j + \Delta x_j(\dot{P}_s^-)_j \tag{7.4.40}$$

The initial guess for $q_{j+\frac{1}{2}}$ will be called $\bar{q}$. It is constructed by computing $\bar{v}$, the arithmetic mean of v_j and v_{j+1}, and then reconstructing $q(\bar{v})$ using (7.2.11) through (7.2.15). This corresponds to $\phi_{j+\frac{1}{2}} = 1$. Using $q_{j+\frac{1}{2}} = \bar{q}$, it is possible to construct $\overline{(\dot{P}_s^+)_j}$ and $\overline{(\dot{P}_s^-)_j}$ by using (7.2.6) and (7.2.7). Summing these gives

$$\overline{(\dot{P}_s)_j} = \overline{(\dot{P}_s^+)_j} + \overline{(\dot{P}_s^-)_j} \tag{7.4.41}$$

As in Chapter 6, a scheme which obeys the constraints

$$(\dot{P}_s^+)_j \geq \min\left(-\overline{(\dot{P}_s^-)_j}, 0\right) \tag{7.4.42}$$

$$(\dot{P}_s^-)_j \geq \min\left(-\overline{(\dot{P}_s^+)_j}, 0\right) \tag{7.4.43}$$

is second-order accurate. The only remaining difficulty is the selection of $q_{j+\frac{1}{2}}$ which satisfies these constraints.

For a particular cell, the value of $(\dot{P}_s^+)_j$ required to satisfy (7.4.42) may be greater than $\overline{(\dot{P}_s^+)_j}$. The difference is defined as

$$\Delta(\dot{P}_s^+)_j \equiv \max\left(-\overline{(\dot{P}_s^-)_j}, \overline{(\dot{P}_s^+)_j}\right) - \overline{(\dot{P}_s^+)_j} \tag{7.4.44}$$

Using (7.4.38) and an assumption of linearity, the change in $\phi_{j+\frac{1}{2}}$ can be computed.

$$\frac{\partial(\dot{P}_s^+)_j}{\partial\phi_{j+\frac{1}{2}}} \cdot \Delta\phi_{j+\frac{1}{2}} = \Delta(\dot{P}_s^+)_j \tag{7.4.45}$$

It is well known that smallest changes in $\phi_{j+\frac{1}{2}}$ (in the sense of length) occur when $\Delta\phi_{j+\frac{1}{2}}$ is chosen along the gradient. That is

$$\Delta\phi_{j+\frac{1}{2}} = \alpha\frac{\partial(\dot{P}_s^+)_j}{\partial\phi_{j+\frac{1}{2}}} \tag{7.4.46}$$

where α is a scalar quantity. When this choice is substituted into (7.4.45) the result is

$$\Delta\phi_{j+\frac{1}{2}} = \Delta(\dot{P}_s^+)_j \frac{\dfrac{\partial(\dot{P}_s^+)_j}{\partial\phi_{j+\frac{1}{2}}}}{\left\| \dfrac{\partial(\dot{P}_s^+)_j}{\partial\phi_{j+\frac{1}{2}}} \right\|^2} \tag{7.4.47}$$

This value is used to compute a new value of $\phi_{j+\frac{1}{2}}$. From $\phi_{j+\frac{1}{2}}$, a new value of $\tilde{v}_{j+\frac{1}{2}}$ is computed using equation (7.4.13). A new value of $v_{j+\frac{1}{2}}$ is computed using (7.4.22). The value of $v_{j+\frac{1}{2}}$ is then used to compute a new value of $q_{j+\frac{1}{2}}$ using (7.2.11) through (7.2.15). Recall that the matrix Y^{-1} is given in Chapter 2 and is evaluated at $q_{j+\frac{1}{2}}$.

There will also be a change in $(\dot{P}_s^-)_{j+1}$ when $q_{j+\frac{1}{2}}$ is updated. This change can be estimated using (7.4.39). It has already been shown that the two gradients $\dfrac{\partial(\dot{P}_s^+)_j}{\partial\phi_{j+\frac{1}{2}}}$ and $\dfrac{\partial(\dot{P}_s^-)_{j+1}}{\partial\phi_{j+\frac{1}{2}}}$ have the same sign, term by term. Consequently we can conclude that their dot product is positive, their included angle less than 90 degrees. Thus, since the change in $\phi_{j+\frac{1}{2}}$ is along $\dfrac{\partial(\dot{P}_s^+)_j}{\partial\phi_{j+\frac{1}{2}}}$, it has a positive component along $\dfrac{\partial(\dot{P}_s^-)_{j+1}}{\partial\phi_{j+\frac{1}{2}}}$. This implies an increase in $(\dot{P}_s^-)_{j+1}$. By implication, because the same procedure is carried out at each cell, $(\dot{P}_s^-)_j$ can be expected not to decrease.

Due to the convexity of the entropy function, the actual change in $(\dot{P}_s^+)_j$ will be less than the linear approximation would suggest. Exceptions occur when roundoff error becomes significant. Since the change falls short of what is needed, iteration is required. The function $(\dot{P}_s^+)_j$ is reevaluated, new gradients are computed and the definition of $\tilde{v}$ changes. A subtlety is that $\phi_{j+\frac{1}{2}}$ changes too, a consequence of the changing definition of $\tilde{v}$. Iteration is continued until $\Delta\phi_{j+\frac{1}{2}}$ is less than some tolerance.

7.5 The Effect of Area Ratio and Metric Terms

In the quasi-one-dimensional Euler equations (which describe flow through a nozzle), the effect of varying cross sectional area appears through a source term in the momentum equation. This corresponds to a pressure acting normal to the nozzle surface. In addition, the area ratio scales the usual fluxes in a manner analogous to metric terms in two-dimensional problems. In this section, the effect of the source term on the cross sectional area is investigated.

A typical differencing scheme for solving the quasi-one-dimensional Euler equations is

$$(q_{,t})_j + \frac{1}{V_j}\left[\left\{(af)_{j+\frac{1}{2}} - (af)_{j-\frac{1}{2}}\right\} - h_j\left\{a_{j+\frac{1}{2}} - a_{j-\frac{1}{2}}\right\}\right] = 0 \tag{7.5.1}$$

where $(af)_{j+\frac{1}{2}} = a_{j+\frac{1}{2}} f_{j+\frac{1}{2}}$. The quantity $a_{j+\frac{1}{2}}$ is the cross-sectional area of the nozzle at $x_{j+\frac{1}{2}}$. The quantity V_j is the volume of the j^{th} cell, typically $V_j = a_j(x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}})$. The vector h is given by

$$h \equiv \begin{bmatrix} 0 \\ p \\ 0 \end{bmatrix} \tag{7.5.2}$$

which is the source term for these equations. Notice that the details of how $f_{j+\frac{1}{2}}$ is computed are not given; the scheme is unspecified except for conservation and the treatment of the metric terms.

The entropy inequality which corresponds to (7.5.1) is

$$(S_{,t})_j + \frac{1}{V_j}\left[(aF)_{j+\frac{1}{2}} - (aF)_{j-\frac{1}{2}}\right] \geq 0 \tag{7.5.3}$$

Using the chain rule on the first term gives

$$(S_{,q})_j (q_{,t})_j + \frac{1}{V_j}\left[(aF)_{j+\frac{1}{2}} - (aF)_{j-\frac{1}{2}}\right] \geq 0 \tag{7.5.4}$$

The factor $(q_{,t})_j$ can be replaced using equation (7.5.1) to give (after multiplying through by the volume)

$$-(S_{,q})_j \left[\left\{(af)_{j+\frac{1}{2}} - (af)_{j-\frac{1}{2}}\right\} - h_j \left\{a_{j+\frac{1}{2}} - a_{j-\frac{1}{2}}\right\}\right] + \left[(aF)_{j+\frac{1}{2}} - (aF)_{j-\frac{1}{2}}\right] \geq 0 \tag{7.5.5}$$

At this point it helps to note the following useful identity which may be verified by construction (perhaps using MACSYMA).

$$S_{,q} h = S_{,q} f - F \tag{7.5.6}$$

This may be applied directly to equation (7.5.5) to yield

$$\begin{aligned} &\left[(S_{,q})_j f_j - F_j\right]\left(a_{j+\frac{1}{2}} - a_{j-\frac{1}{2}}\right) \\ &\quad -(S_{,q})_j \left((af)_{j+\frac{1}{2}} - (af)_{j-\frac{1}{2}}\right) + \left((aF)_{j+\frac{1}{2}} - (aF)_{j-\frac{1}{2}}\right) \geq 0 \end{aligned} \tag{7.5.7}$$

Applying the definitions of af and aF and grouping terms gives the final form.

$$\begin{aligned} &a_{j+\frac{1}{2}}\left[-(S_{,q})_j (f_{j+\frac{1}{2}} - f_j) + (F_{j+\frac{1}{2}} - F_j)\right] \\ &+ a_{j-\frac{1}{2}}\left[-(S_{,q})_j (f_j - f_{j-\frac{1}{2}}) + (F_j - F_{j-\frac{1}{2}})\right] \geq 0 \end{aligned} \tag{7.5.8}$$

This form is independent of how the fluxes are computed and how the cell volume is computed. The source term does not appear. This agrees with classical thermodynamics; the source term is a pure force and therefore does not contribute to the entropy production rate.

7.6 Results and Discussion

For the test problem, a converging, diverging nozzle geometry was used. The geometry is described by the formula

$$a(x) \equiv 0.8 * \left[1 + \left(x - \frac{L}{2}\right)^2\right] \tag{7.6.1}$$

Where L is the length of the nozzle; in this case $L = 1$. The geometry is shown in figure 7.1.

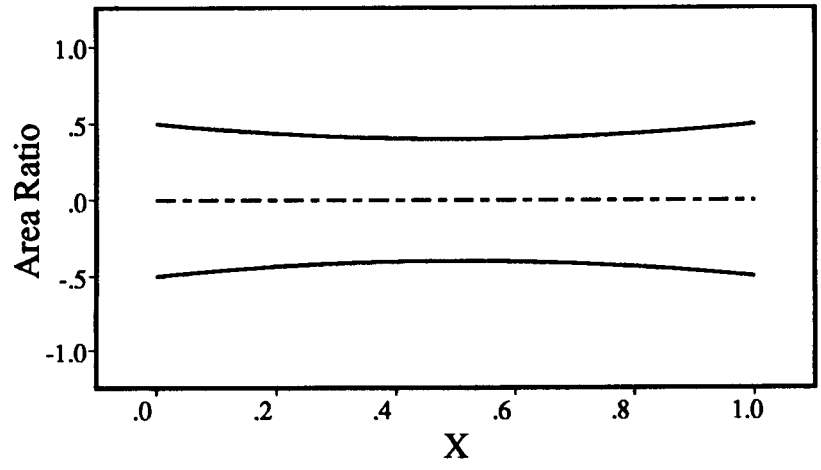

Figure 7.1. – Nozzle geometry used for a test case. The conditions used had subsonic inflow and outflow with a supersonic region between $x = 0.5$ and $x = 0.8$.

The particular case chosen has a steady solution with subsonic inflow and outflow. The flow is supersonic between the throat ($x = 0.5$) and a shock which is located at $x = 0.8$.

The first test was to verify that the semi-discrete scheme described in Section 7.4 worked correctly. The steady state exact solution was used to test the scheme. The effectiveness of the gradient following technique can be seen in figure 7.2. The solid line represents $\overline{(\dot{P}_s)_j}$, which uses linear averages of the entropy variables to compute $q_{j+\frac{1}{2}}$. This results in a negative entropy production rate over much of the domain. In order to use the stability conjecture of Chapter 2, a positive entropy production rate for each cell was sought. The diamonds represent the minimum entropy production rate guaranteed by constraints (7.4.42) and (7.4.43). The dashed line, which represents $(\dot{P}_s)_j$, shows what was actually achieved when (7.4.42) and (7.4.43) were met.

The quantity $q_{,t}$, computed by the semi-discrete scheme, found its largest values at the sonic point ($x = 0.6$) and at the shock ($x = 0.8$). The entropy production rate at the

shock was four orders of magnitude higher than any other point in the domain. This is to be expected from the physics. The near-zero entropy production rate at the sonic point is an artifact of the method; at sonic points, the constraints (7.4.20) and (7.4.21) translate to $(\dot{P}_s^+)_j = (\dot{P}_s^-)_j = (\dot{P}_s)_j = 0$.

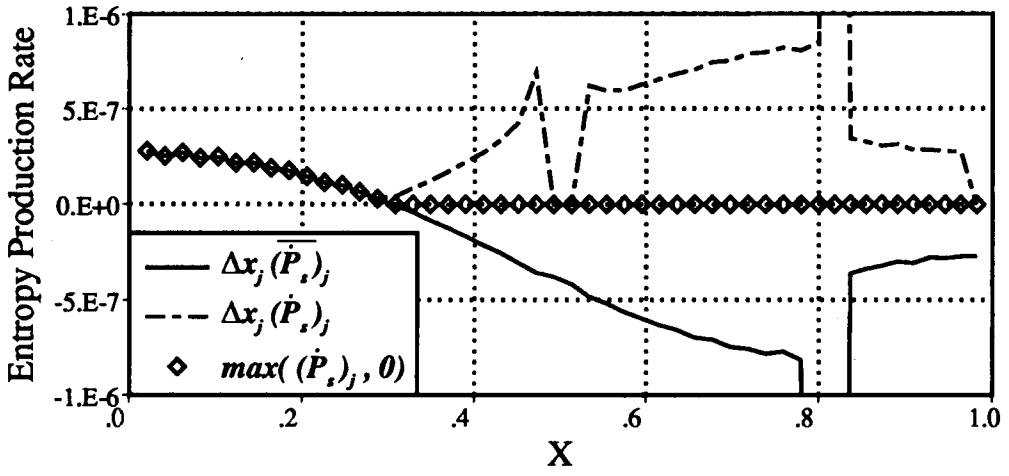

Figure 7.2. – Semi-discrete entropy production rates for the exact solution. The solid line comes from using $v_{j+\frac{1}{2}} = \bar{v}$. The chaindash line comes from modifying $v_{j+\frac{1}{2}}$ to satisfy a cell entropy inequality.

In Chapter 4, the question of time advance was explored. It was pointed out for example that explicit Euler is not a suitable time advance scheme for convective problems such as this one. The fully discrete entropy production rate for this scheme will fall short of the semi-discrete entropy production rate by an amount that is proportional to the change in the solution. This in turn is proportional to the time step. This conclusion was tested numerically, again using the exact solution for test data. In figure 7.3, the difference between the semi-discrete and fully discrete production rates is shown for various values of the time step. Notice that the loss in entropy production rate is considerably greater than the rate itself (fig. 7.2), especially around the sonic point. This would imply negative entropy production rates for the fully discrete scheme. The almost perfect proportionality of the curves (note the log scale) indicates that the uncertainty in the evaluation time of $S_{,qq}$ (4.2.8) makes little difference. It might as well have been evaluated at q^n. The flattening of the curves at very low values is due to roundoff errors.

In this work, implicit Euler was employed. In contrast to explicit Euler, it has the property that the fully discrete entropy production rate will always be greater than the semi-discrete entropy production rate. In principle this implies unconditional, nonlinear stability. For example, in figure 7.4, the entropy production due to time advance is shown

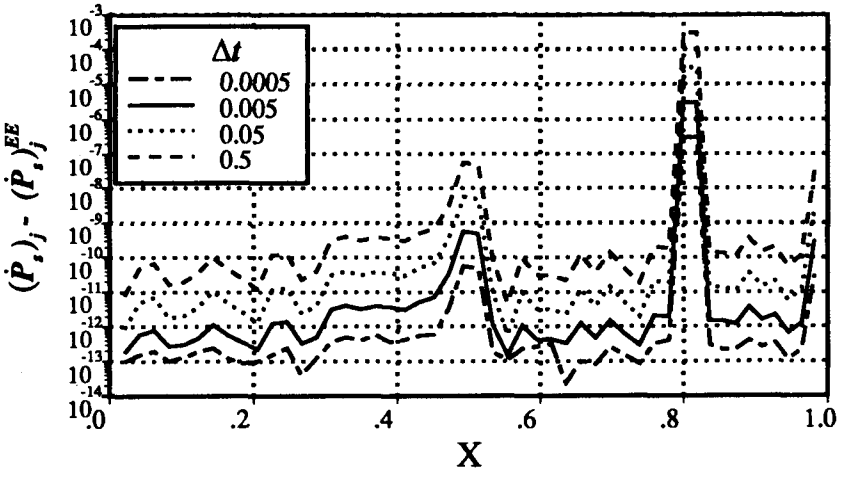

Figure 7.3. – Explicit Euler time advance destroys entropy. As Chapter 4 pointed out, the amount is roughly proportional to Δt.

as a function of the time step. The entropy production rate increases as the time step increases but not proportionately. This is due to differences in the evaluation state of $S_{,qq}$ (4.3.8). This state is approximately q^{n+1}, which is different for each of the curves.

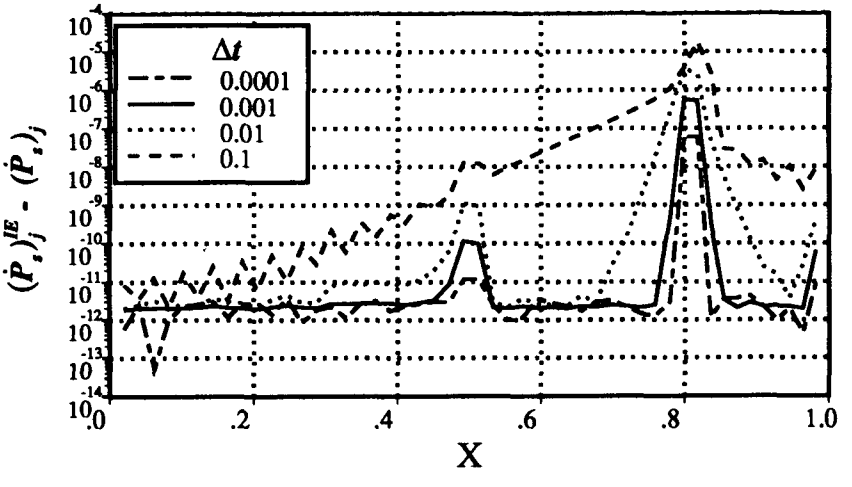

Figure 7.4. – Implicit Euler time advance creates entropy. This effect is especially pronounced at larger time steps.

In practice, there are several implementation difficulties, not all of which were discussed in Section 4.6. One difficulty is that to obtain all the benefits of an implicit scheme requires solution of a system of algebraic equations that become increasingly nonlinear at larger values of Δt. Essentially this requires evaluation of the fluxes at a future time. Since any value of Δt can be used, this is equivalent to knowing the time-dependent solution

C-2

in advance. Such knowledge is not generally available; indeed it would obviate the need for numerical methods. As far as nonlinear stability is concerned, implicit methods are restricted to time steps sufficiently small that linear approximations in q are valid. For example, the smallest time steps used in figure 7.4 were constrained by roundoff error, while the largest time steps were limited by the ability to converge the iterative process by which q^{n+1} was obtained.

In practice, terms involving q^{n+1} are linearized about the current solution using a Taylor series. For example

$$f_{j+\frac{1}{2}}^{n+1} = f_{j+\frac{1}{2}}^{n} + \sum_i \frac{\partial f_{j+\frac{1}{2}}}{\partial q_i}(q_i^{n+1} - q_i^n) \tag{7.4.7}$$

The sum is taken over each unknown in the whole mesh, but many of the terms vanish. For this scheme, $f_{j+\frac{1}{2}}$ depends only on $q_{j-1}, q_j, q_{j+1}, q_{j+2}$. Since analytical expressions for the derivatives were difficult to obtain (to say the least), it was decided to evaluate the derivatives in (7.4.7) numerically by perturbing q slightly and measuring the the changes in f which result. This requires twelve additional evaluations of $q_{j+\frac{1}{2}}$ per step (the number of variables per node times the width of the stencil).

Once the derivatives are computed, a standard penta-diagonal matrix solver is used to advance in time. This gives an approximate value for q^{n+1}. For computing figure 7.4, this process was iterated until a very accurate value of q^{n+1} was obtained. On the other hand, for the main computation, the modification discussed in Section 4.5 was used to allow an accurate measure of the fully discrete entropy production rate. The modification has no effect at all when f is a linear function of q.

It was noted with dismay that linearization in time has a disastrous effect on stability. For small time steps, the linearization is a good approximation to the nonlinear reality and the predicted behavior occurs. For Courant numbers greater than one, however, the linearization errors may be sufficient to cause local violations of the second law. These errors were largest during transients, especially those with moving shocks. As convergence approached, it became possible to take slightly larger time steps. With exact updates, in principle, there is no linearization error and no time step limitation.

The last test consisted of starting with a fairly arbitrary initial condition and allowing it to evolve in time to reach a steady state. The initial condition chosen was obtained by linear interpolation of the conserved variables between the endpoints. A constant time step of 0.05 was used, roughly equivalent to a CFL number of $\frac{1}{2}$. The initial entropy production rate distribution is shown in figure 7.5.

The solid line shows the fully discrete entropy production rate. Notice that it is greater than the semi-discrete entropy production rate (represented by the diamonds)

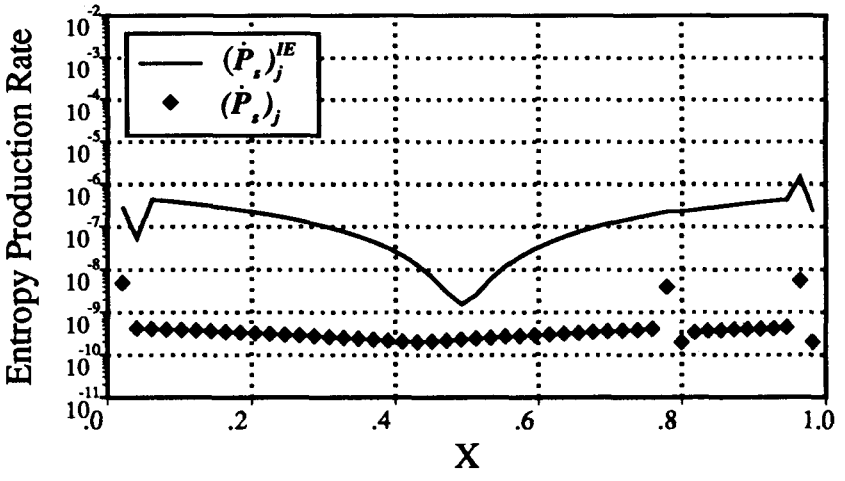

Figure 7.5. – Initial entropy production rate. The initial conditions are a simple linear interpolation of the boundary conditions.

at each grid point. The difference is proportional to the change in q in the first time step. There is a slight glitch in the semi-discrete entropy production rate at $x = 0.8$, the eventual location of the shock. This glitch does not reflect numerical prescience, but rather a very slight metric discontinuity, an artifact of the mesh generation scheme (the mesh was almost equally spaced, but one point was placed at the shock location). This tiny glitch doesn't appear in the fully discrete entropy production rate; it is swamped by the temporal dissipation (note the three order of magnitude difference in the two rates). Similar glitches at the boundaries are simply initial transients. To avoid roundoff problems, a small amount of entropy is intentionally produced in each cell. This leads to an artificial floor of 10^{-10} on the entropy production rate.

The exact solution was used to specify the conservative variables at inflow and outflow. This formally overspecifies the boundary conditions, so minor glitches at the boundaries can be expected.

The initial condition was advanced to a nondimensional time of 25, at which time the solution appeared to be converged. The plot of density is shown in figure 7.6. The exact solution is shown as a solid line, and the computed solution is shown as the diamonds. Also plotted are the values of $q_{j+\frac{1}{2}}$ from which the flux is computed (vertical ticks). Notice that none of these appear "in" the shock.

Although the energy appears much the same as the density (and for that reason isn't shown here) the momentum looks substantially different. It is shown in figure 7.7. Notice the substantial oscillation in the computed solution around the shock location. Although the computed solution oscillates, the values of $q_{j+\frac{1}{2}}$ are very close to the exact solution.

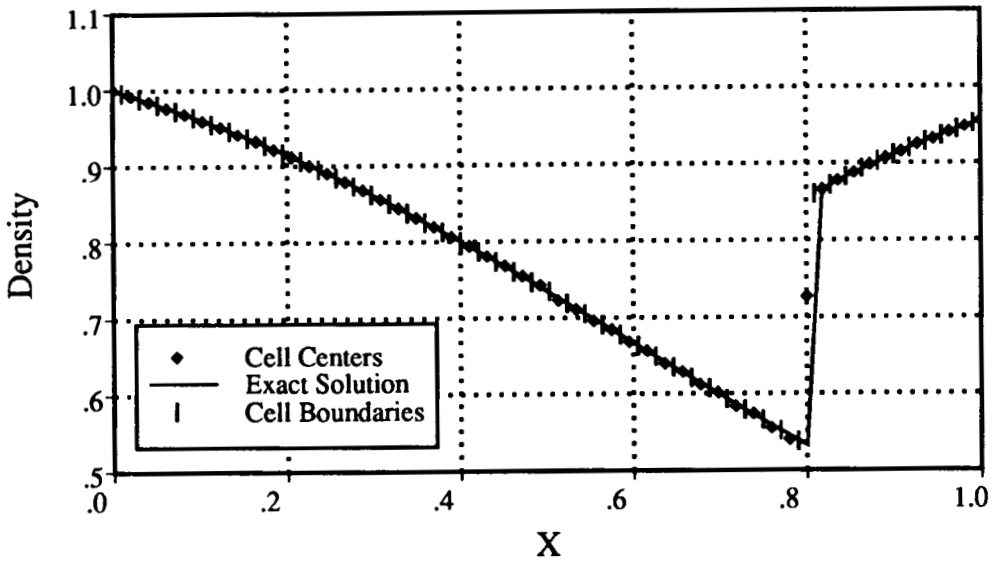

Figure 7.6. – Final density distribution.

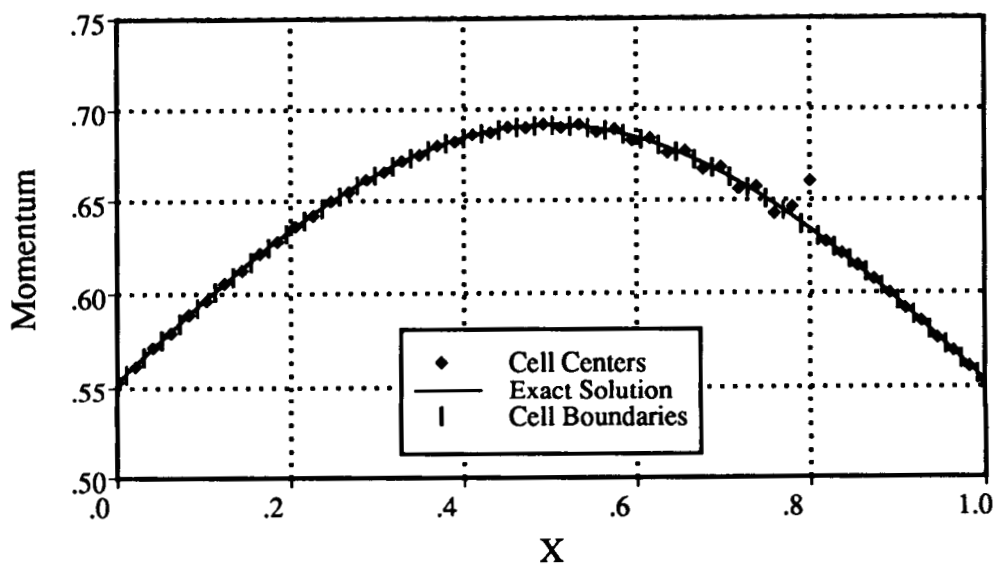

Figure 7.7. – Final momentum distribution. The computed solution matches the exact solution well on cell boundaries, where fluxes are balanced. It matches less well at cell centers, especially at the shock. Even though an interpolation is monotone in the entropy variables, it may not be monotone in the conservative variables.

This brings up an interesting question concerning uniqueness of the solution. The solution is considered converged when it stops changing, i.e. when the fluxes reach their exact values. It seems likely that more than one solution is capable of interpolating the same values of $q_{j+\frac{1}{2}}$. Thus, there may be more than one solution.

The interpolation is limited by the consistency condition, which is applied to a particular set of characteristic variables. Enforcing an interval boundedness condition in these variables does not imply such a condition in other variables, due to the nonlinearity of the transformation. The solution shown in figure 7.7 is a case in point; there is an interval immediately to the left of $x = 0.8$ for which $(\rho ua)_{j+\frac{1}{2}}$ does not fall between $(\rho ua)_j$ and $(\rho ua)_{j+1}$.

Other researchers[2,3] have suggested that one cannot expect interval boundedness in momentum at a shock on a finite mesh. The uncertainty in the shock location (it is only known to be somewhere in the interval) permits the computed state within the shock to vary within some limits. Its steady state value is a function of the initial conditions. The interval bounded quantities are ρ, u, and p. Under these conditions, interval boundedness for momentum occurs only in the degenerate case where the shock is reduced to a single interval. Ultimately, the problem stems from the approximation that q_j is equal to the cell average. All schemes which use this approximation (including Godunov's scheme and others) generate a similar glitch in the solution.

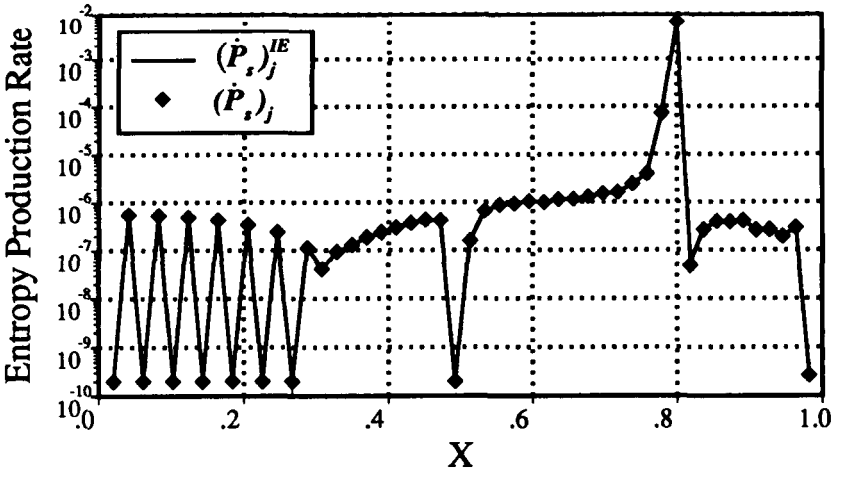

Figure 7.8. – Final entropy production rate. Notice the strong peak at the shock ($x = 0.8$), as required by physics. The dip at the sonic point ($x = 0.5$) is an artifact of the numerical scheme. The low amplitude oscillations at the left boundary may be due to the overspecified boundary conditions. The agreement between semi-discrete and fully discrete values indicates very small changes in q.

The final entropy production rate is shown in figure 7.8. Notice that the semi-discrete and fully discrete production rates correspond, as they must at convergence. Notice the four-order-of-magnitude spike in entropy production rate which occurs at the shock

[2] T. Barth, personal communications

[3] T. Barth, Some Notes on Shock Resolving Flux Functions, 1988 (unpublished)

($x = 0.8$). The analytic solution shows a Dirac spike at this location, accounting for all the entropy produced in the domain. There is no oscillation in the computed entropy production rate near the shock to correspond with the observed oscillation in momentum. This is because steady state entropy production rates depend only on $q_{j+\frac{1}{2}}$ values (the tiks in figure 7.7) which exhibit no such oscillations. Another feature of interest is a dramatic dip in entropy production around the sonic point ($x = 0.5$), where the constraints (7.4.20) and (7.4.21) evaluate to $(\dot{P}_s^+)_j = (\dot{P}_s^-)_j = 0$. Finally, some ringing in the entropy production rate is apparent at the inflow boundary. This is probably due to a slight inconsistency in the boundary conditions, which are overspecified. There is no apparent effect on the solution from this ringing.

Overall then, it appears that satisfaction of a cell entropy inequality is sufficient to produce a stable scheme. Such a scheme has very little dissipation. On the bad side, it evidently isn't sufficient to avoid the unphysical oscillations in momentum and it has a severe time step limitation.

Despite its practical limitations, this scheme is interesting for its theoretical properties. Its computation of actual values for $q_{j+\frac{1}{2}}$ is shared only by Godunov's scheme, a first-order explicit scheme. The second law allows difficulties with grid spacing or time step to be pinpointed. This approach also provides insight to grid embedding, multigrid, and other interpolation-related problems.

Chapter 8. THE TVD CONNECTION

8.1 Introduction

This chapter reviews the basic concept of Total Variation Diminishing (TVD) schemes. Where applicable, it explores the connections between the TVD approach and the satisfaction of a cell entropy inequality. Many of the TVD ideas shown here are due to Harten (ref. 11). They are also found in an excellent survey paper by Sweby (ref. 27).

Harten observed that existing schemes for solving convection problems were unsatisfactory with regard to dissipation. One choice was monotone schemes (ref. 28). These schemes are guaranteed not to produce any new extrema. This property is useful in preventing negative densities for example, in computations involving very strong shocks. Unfortunately, all linear monotone schemes are at best first-order accurate spatially (ref. 29).

The other major choice consisted of adding constant coefficient dissipative terms to the PDEs as suggested by Lax (ref. 1). Since these are not part of the original equations, the coefficient is made as small as possible. Straightforward, symmetric differencing of the dissipative terms leads to stable schemes with second-order accuracy. An unfortunate side effect is the introduction of new extrema, oscillations, especially in the neighborhood of shocks.

Thus algorithm designers were faced with a choice between first order methods which are robust, and higher order methods which aren't. In either case, the dissipation is linear, easy to apply, and easy to analyze using Fourier analysis.

A number of ad hoc methods emerged to make the dissipation coefficients dependent on the local solution. For example, Jameson suggested that the dissipation should consist of a linear second difference term multiplied by a term which depends on the second difference of pressure (ref. 30). This "sensed" the presence of shocks and locally increased the dissipation there to remove oscillations. Pulliam (ref. 31) showed a qualitative similarity between this approach and upwind differencing.

Another approach was taken by Boris and Book in their Flux Corrected Transport (FCT) approach (ref. 10). A second-order accurate scheme was used to compute fluxes between cells. Where these would result in new extrema, a first-order, monotone flux was also computed and an interpolation between the two was used. The interpolation was chosen in such a way as to eliminate the undesirable new extrema.

This strategy was formalized by Harten in his TVD schemes (ref. 11). Harten abandoned the requirement of monotonicity in order to achieve second order accuracy. Instead, he substituted the requirement that the Total Variation (TV) of the solution must not

increase with time. The concept of total variation is formally and unambiguously defined only for scalar equations in one space dimension. Given a solution vector u defined on J grid points, TV is defined as

$$TV(u) = \sum_{j=1}^{J} |u_{j+1} - u_j| \tag{8.1.1}$$

A TVD scheme is one in which this quantity can be relied on not to grow in time. Generally, periodic boundary conditions are assumed in the analysis. Very recently there has been some work on finite domain TVD schemes (ref. 32), but this is not discussed here.

Harten explained how to construct schemes which are TVD. In particular, if the difference scheme chosen can be expressed in the form

$$u_j^{n+1} = u_j^n - C_{j-\frac{1}{2}}(u_j^n - u_{j-1}^n) + D_{j+\frac{1}{2}}(u_{j+1}^n - u_j^n) \tag{8.1.2}$$

where $C_{j-\frac{1}{2}}$ and $D_{j+\frac{1}{2}}$ are data-dependent scalar coefficients, and if

$$\begin{aligned} C_{j-\frac{1}{2}} &\geq 0 \\ D_{j+\frac{1}{2}} &\geq 0 \\ C_{j+\frac{1}{2}} + D_{j+\frac{1}{2}} &\leq 1 \end{aligned} \tag{8.1.3}$$

then the scheme is TVD. The proof of this assertion is subtle but short. It is given in Appendix A in some detail. Conditions (8.1.3) are sufficient, but not necessary; there may be TVD schemes which do not meet these conditions.

The concept of TVD schemes has been extended to implicit methods, through the use of additional conditions. It extends to multidimensions on regular grids through the use of operator splitting. Finally it extends to certain systems of equations through the use of eigensystem decomposition. A reasonable summary can be found in the work of Yee (ref. 33).

The goal of the TVD approach is to produce solutions which are aesthetically pleasing. This is not necessarily the same as solutions that are correct (see fig. 5.4). The second law of thermodynamics provides a way to tell the two apart.

8.2 Scalar Wave Equation

This section explores the application of TVD ideas to the scalar wave equation. It also spells out the relationship between the TVD conditions and the requirements for a cell entropy inequality.

Consider the scalar wave equation:

$$u_{,t} + cu_{,x} = 0, \quad c > 0 \tag{8.2.1}$$

This is typically approximated by a central difference with a dissipation term

$$u_j^{n+1} = u_j^n - \frac{1}{2}\nu(u_{j+1}^n - u_{j-1}^n) + \frac{1}{2}\nu[\alpha_{j+\frac{1}{2}}(u_{j+1}^n - u_j^n) - \alpha_{j-\frac{1}{2}}(u_j^n - u_{j-1}^n)] \tag{8.2.2}$$

The parameter $\alpha_{j+\frac{1}{2}}$ is a dissipation coefficient which may or may not vary from point to point. As written, equation (8.2.2) describes a conservative scheme which is consistent as long as α is less than unity and stable when α is positive. The parameter ν is the CFL number: $\nu \equiv \frac{c\Delta t}{\Delta x}$.

At this point, (8.2.2) must be manipulated to look more like (8.1.2) so that the TVD conditions can be applied. The notation is simplified if u_{j+1}^n and u_{j-1}^n are expressed in terms of differences.

$$\begin{aligned} \Delta u_{j+\frac{1}{2}} &\equiv u_{j+1}^n - u_j^n \\ \Delta u_{j-\frac{1}{2}} &\equiv u_j^n - u_{j-1}^n \end{aligned} \tag{8.2.3}$$

With these definitions (8.2.2) becomes

$$u_j^{n+1} = u_j^n - \frac{1}{2}\nu(\Delta u_{j+\frac{1}{2}} + \Delta u_{j-\frac{1}{2}}) + \frac{1}{2}\nu(\alpha_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}} - \alpha_{j-\frac{1}{2}}\Delta u_{j-\frac{1}{2}}) \tag{8.2.4}$$

By combining terms, this becomes

$$u_j^{n+1} = u_j^n - \frac{1}{2}\nu(1 - \alpha_{j+\frac{1}{2}})\Delta u_{j+\frac{1}{2}} - \frac{1}{2}\nu(1 + \alpha_{j-\frac{1}{2}})\Delta u_{j-\frac{1}{2}} \tag{8.2.5}$$

Sweby (ref. 27) defines $\phi_{j+\frac{1}{2}} \equiv 1 - \alpha_{j+\frac{1}{2}}$. With this definition, (8.2.5) becomes

$$u_j^{n+1} = u_j^n - \frac{1}{2}\nu\phi_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}} - \frac{1}{2}\nu(2 - \phi_{j-\frac{1}{2}})\Delta u_{j-\frac{1}{2}} \tag{8.2.6}$$

By inspection, the only constant value of ϕ which fits the TVD conditions (8.1.2) is $\phi = 0$. This results in the familiar first-order upwind scheme. There is, however an elegant trick which improves accuracy. Sweby defines the quantity $r_j \equiv \frac{\Delta u_{j-\frac{1}{2}}}{\Delta u_{j+\frac{1}{2}}}$, and eliminates $\Delta u_{j+\frac{1}{2}}$ from the equation.

$$u_j^{n+1} = u_j^n - \nu\left[1 + \frac{1}{2}\left(\frac{\phi_{j+\frac{1}{2}}}{r_j} - \phi_{j-\frac{1}{2}}\right)\right]\Delta u_{j-\frac{1}{2}} \tag{8.2.7}$$

which also has the the form of (8.1.2). The scheme is TVD if

$$0 \le \nu \left[1 + \frac{1}{2} \left(\frac{\phi_{j+\frac{1}{2}}}{r_j} - \phi_{j-\frac{1}{2}} \right) \right] \le 1 \tag{8.2.8}$$

This is equivalent, for positive ν to

$$-2 \le \frac{\phi_{j+\frac{1}{2}}}{r_j} - \phi_{j-\frac{1}{2}} \le \frac{2(1-\nu)}{\nu} \tag{8.2.9}$$

For $\nu \le \frac{1}{2}$, this becomes

$$-2 \le \frac{\phi_{j+\frac{1}{2}}}{r_j} - \phi_{j-\frac{1}{2}} \le 2 \tag{8.2.10}$$

This can be separated further into the inequalities

$$0 \le \phi_{j+\frac{1}{2}} \le 2 \tag{8.2.11}$$

$$0 \le \frac{\phi_{j+\frac{1}{2}}}{r_j} \le 2 \tag{8.2.12}$$

These inequalities require that $\phi_{j+\frac{1}{2}} = 0$ if $r_j \le 0$. For $r_j \ge 0$, the inequalities are plotted in figure 8.1. As long as $\phi_{j+\frac{1}{2}}(r_j)$ lies within the shaded region, both inequalities will hold and the scheme will enjoy the TVD property. To achieve second-order accuracy in the limit as $\Delta x \to 0$ requires only that $\phi_{j+\frac{1}{2}}(1) = 1$ and that $\phi_{j+\frac{1}{2}}$ be a continuous function of r_j. (Several second-order accurate schemes are shown in figure 5.3)

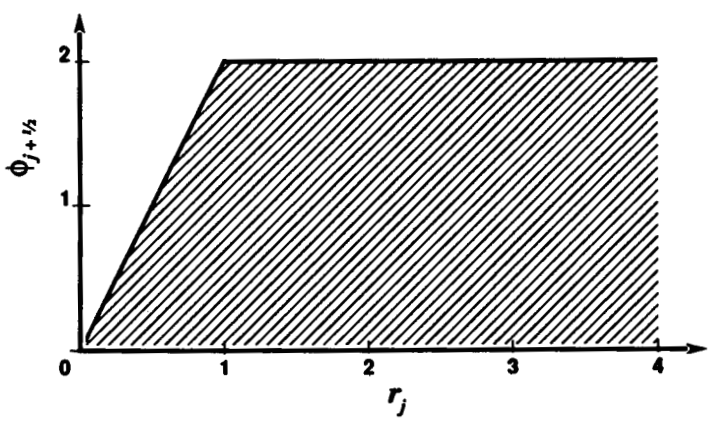

Figure 8.1. – Schemes in the shaded region satisfy the TVD conditions (8.1.3). See figure 5.3 for some typical schemes.

There is a connection between TVD ideas and the entropy analysis discussed in previous sections. For the wave equation the connection is known precisely. In Chapter 5, the semi-discrete entropy analysis was shown for this problem (eq. (5.1.8)). What is needed here is a fully discrete analysis using explicit Euler time advance. This is derived in Section 5.4 (5.4.6) and is restated here.

$$\Delta t(\dot{P}_s)_j^{EE} = \nu\left[-(u^n_{j+\frac{1}{2}} - u^n_j)^2 + (u^n_j - u^n_{j-\frac{1}{2}})^2\right] - \nu^2(u^n_{j+\frac{1}{2}} - u^n_{j-\frac{1}{2}})^2 \tag{8.2.13}$$

If ν is positive (the case where ν is negative is a trivial extension), then a sufficient condition for positive entropy production rates is (5.4.9), plotted in figure 5.2. This is restated here for convenience.

$$-1 \le \frac{u^n_{j+\frac{1}{2}} - u^n_j}{u^n_j - u^n_{j-\frac{1}{2}}} \le \frac{1-\nu}{1+\nu} \tag{8.2.14}$$

It remains to define $u_{j+\frac{1}{2}}$ and $u_{j-\frac{1}{2}}$. An elegant choice is (6.2.5)

$$u_{j+\frac{1}{2}} = u_j + \frac{1}{2}\phi_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}} \tag{8.2.15}$$

and an analogous definition for $u_{j-\frac{1}{2}}$. With these definitions and the definition of r_j previously given, inequalities (8.2.14) simplify to

$$-1 \le \frac{\phi_{j+\frac{1}{2}}/r_j}{2 - \phi_{j-\frac{1}{2}}} \le \frac{1-\nu}{1+\nu} \tag{8.2.16}$$

These inequalities (for $\nu = \frac{1}{2}$) are plotted as the shaded region of figure 8.2. From previous analysis, the region between the two parallel lines satisfies inequalities which meet the TVD conditions (8.2.10). The area to the right of $\phi = 2$ is generally excluded on consistency grounds. Thus the cell entropy inequality is more restrictive than the TVD conditions in this case. Schemes which are consistent and satisfy the second law are TVD schemes. The converse is not true.

8.3 Burgers' Equation

In this section, a TVD scheme is derived for Burgers' equation. A nonstandard approach is used to facilitate comparison with the scheme derived in Section 6.3. A more standard approach will be demonstrated in the next section.

Consider a generic conservative scheme for solving Burgers' equation.

$$u^{n+1}_j = u^n_j - \frac{\Delta t}{\Delta x}(f^n_{j+\frac{1}{2}} - f^n_{j-\frac{1}{2}}) \tag{8.3.1}$$

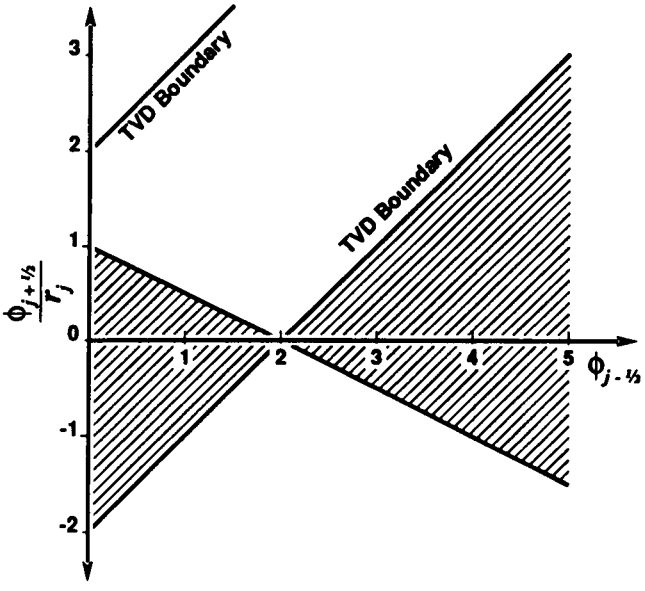

Figure 8.2. – Schemes between the parallel lines satisfy the TVD constraints. Schemes in the shaded region satisfy a cell entropy inequality. Schemes where $\phi_{j+\frac{1}{2}} > 2$ are generally not used. Thus the entropy inequality is more restrictive than the TVD constraints.

For Burgers' equation, this reduces to

$$u_j^{n+1} = u_j^n - \frac{1}{2}\frac{\Delta t}{\Delta x}\left[(u_{j+\frac{1}{2}}^n)^2 - (u_{j-\frac{1}{2}}^n)^2\right] \tag{8.3.2}$$

Factoring the difference of two squares gives

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}\frac{1}{2}(u_{j+\frac{1}{2}}^n + u_{j-\frac{1}{2}}^n)(u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n) \tag{8.3.3}$$

This looks like the wave equation except that the wave speed is

$$c_j = \frac{1}{2}(u_{j+\frac{1}{2}}^n + u_{j-\frac{1}{2}}^n) \tag{8.3.4}$$

Defining $\nu_j \equiv \frac{c_j \Delta t}{\Delta x}$ gives

$$u_j^{n+1} = u_j^n - \nu_j(u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n) \tag{8.3.5}$$

As before, the definitions of $u_{j+\frac{1}{2}}^n$ and $u_{j-\frac{1}{2}}^n$ are

$$u_{j+\frac{1}{2}}^n = u_j^n + \frac{1}{2}\phi_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}} \tag{8.3.6}$$

$$u_{j-\frac{1}{2}}^n = u_j^n - \frac{1}{2}(2 - \phi_{j-\frac{1}{2}})\Delta u_{j-\frac{1}{2}} \tag{8.3.7}$$

With these definitions (8.3.5) becomes

$$u_j^{n+1} = u_j^n - \frac{1}{2}\nu_j\phi_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}} - \frac{1}{2}\nu_j(2-\phi_{j-\frac{1}{2}})\Delta u_{j-\frac{1}{2}} \tag{8.3.8}$$

which is virtually the same as (8.2.5) except for the definition of ν_j. Thus the analysis of the previous section applies. If ν_j is positive, the scheme is TVD if

$$-2 \le \frac{\phi_{j+\frac{1}{2}}}{r_j} - \phi_{j-\frac{1}{2}} \le \frac{2(1-\nu_j)}{\nu_j} \tag{8.3.9}$$

Recall that in Section 6.4, $(u_{j+\frac{1}{2}} - u_j)$ was assumed to be much less than u_j (leading to (6.4.8)). Here, as there, that assumption results in the approximation that $\nu_j \approx u_j\frac{\Delta t}{\Delta x}$. Under this approximation, the scheme defined in Section 6.3 is identical to the TVD scheme just derived. Thus the connection between schemes which diminish total variation and those which produce entropy depends on the approximation of locally constant wave speed. Section 6.5 discusses the merits and drawbacks of such an approximation.

8.4 Roe's Approach

There is another elegant way of deriving TVD schemes for Burgers' equation. This approach, due to Roe, has the advantage of avoiding the approximation for ν_j. More importantly it is easily (though not rigorously) extended to systems of PDEs.

The approach taken by Roe is to use the intermediate value theorem to compute $f_{j+\frac{1}{2}}$. That is

$$f_{j+1} - f_j = (f_{,u})_{j+\frac{1}{2}}(u_{j+1} - u_j) \tag{8.4.1}$$

where $(f_{,u})_{j+\frac{1}{2}}$ is evaluated at some value of u between u_{j+1} and u_j. The intermediate value theorem guarantees that such a point exists. In particular, for Burgers' equation, it is the linear average of u_j and u_{j+1}.

Substituting the definition of $f(u)$, the left-hand side of (8.4.1) factors.

$$\frac{1}{2}\left[(u_{j+1})^2 - (u_j{}^2)\right] = \frac{1}{2}(u_{j+1}+u_j)(u_{j+1}-u_j) \tag{8.4.2}$$

Equation (8.4.1) holds iff $(f_{,u})_{j+\frac{1}{2}} = \frac{1}{2}(u_{j+1}+u_j)$. Since $f_{,u} = u$, $(f_{,u})_{j+\frac{1}{2}}$ should be evaluated at the linear average of u_j and u_{j+1}.

All of this can be incorporated into a TVD scheme as follows. Consider a central difference scheme with dissipation added, analogous to (8.2.2).

$$\begin{aligned} u_j^{n+1} = u_j^n &- \frac{1}{2}\frac{\Delta t}{\Delta x}(f_{j+1}^n - f_{j-1}^n) \\ &+ \frac{\Delta t}{\Delta x}\left[(f_{,u})_{j+\frac{1}{2}}\alpha_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}} - (f_{,u})_{j-\frac{1}{2}}\alpha_{j-\frac{1}{2}}\Delta u_{j-\frac{1}{2}}\right] \end{aligned} \tag{8.4.3}$$

Manipulating the central difference term gives

$$u_j^{n+1} = u_j^n - \frac{1}{2}\frac{\Delta t}{\Delta x}\left[(f_{j+1}^n - f_j^n) + (f_j^n - f_{j-1}^n)\right] + \frac{\Delta t}{\Delta x}\left[(f_{,u})_{j+\frac{1}{2}}\alpha_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}} - (f_{,u})_{j-\frac{1}{2}}\alpha_{j-\frac{1}{2}}\Delta u_{j-\frac{1}{2}}\right] \tag{8.4.4}$$

Applying (8.4.1) directly, gives

$$u_j^{n+1} = u_j^n - \frac{1}{2}\frac{\Delta t}{\Delta x}\left[(f_{,u})_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}} + (f_{,u})_{j-\frac{1}{2}}\Delta u_{j-\frac{1}{2}}\right] + \frac{\Delta t}{\Delta x}\left[(f_{,u})_{j+\frac{1}{2}}\alpha_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}} - (f_{,u})_{j-\frac{1}{2}}\alpha_{j-\frac{1}{2}}\Delta u_{j-\frac{1}{2}}\right] \tag{8.4.5}$$

At this point the definition $\nu_{j+\frac{1}{2}} \equiv \frac{\Delta t}{\Delta x}(f_{,u})_{j+\frac{1}{2}}$ is introduced. Collecting terms gives.

$$u_j^{n+1} = u_j^n - (\alpha_{j+\frac{1}{2}} + \frac{1}{2})\nu_{j-\frac{1}{2}}\Delta u_{j-\frac{1}{2}} + (\alpha_{j+\frac{1}{2}} - \frac{1}{2})\nu_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}} \tag{8.4.6}$$

Finally, the substitution is made to express α in terms of ϕ as for the wave equation

$$u_j^{n+1} = u_j^n - \frac{1}{2}(2 - \phi_{j-\frac{1}{2}})\nu_{j-\frac{1}{2}}\Delta u_{j-\frac{1}{2}} - \frac{1}{2}\phi_{j+\frac{1}{2}}\nu_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}} \tag{8.4.7}$$

At this point an appropriate definition of r_j is

$$r_j \equiv \frac{\nu_{j-\frac{1}{2}}\Delta u_{j-\frac{1}{2}}}{\nu_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}}} \tag{8.4.8}$$

This can be used to eliminate $\nu_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}}$ from (8.4.7), leading to

$$u_j^{n+1} = u_j^n - \nu_{j-\frac{1}{2}}\left[1 + \frac{1}{2}\left(\frac{\phi_{j+\frac{1}{2}}}{r_j} - \phi_{j-\frac{1}{2}}\right)\right]\Delta u_{j-\frac{1}{2}} \tag{8.4.9}$$

Provided that $\nu_{j-\frac{1}{2}}$ and $\nu_{j+\frac{1}{2}}$ are both positive, this is in the same form as (8.2.7), from which a TVD scheme was constructed. A similar expression holds if both $\nu_{j-\frac{1}{2}}$ and $\nu_{j+\frac{1}{2}}$ are negative. At shocks and sonic points ($r_j < 0$) it is common to drop to first-order accuracy (i.e. $\phi = 0$). Unlike the scheme of Section 8.3, it is possible to evaluate $\nu_{j+\frac{1}{2}}$ exactly, without iteration.

The difference between Burgers' equation and the linear wave equation is the treatment of the nonconstant wave speed. Not surprisingly, this is also one of the main differences between the various TVD and entropy producing schemes discussed so far. The other is the way that sign changes in the wave speed are handled.

8.5 One-Dimensional Gasdynamics, Roe's Identity

In the previous section, the intermediate value theorem was used to factor a flux difference into an intermediate wave speed times a difference in the dependent variables. This approach generalizes to systems of equations as well. That is, there may exist a value of $q_{j+\frac{1}{2}}$ such that

$$f_{j+1} - f_j = (f_{,q})_{j+\frac{1}{2}}(q_{j+1} - q_j) \tag{8.5.1}$$

where $(f_{,q})_{j+\frac{1}{2}}$ means $f_{,q}$ evaluated at $q_{j+\frac{1}{2}}$. As shown by Harten, Lax, and van Leer (ref. 34), such a value exists if an entropy pair (S, F) exists. Still, it may not be easy to find. Fortunately Roe has provided us with $q_{j+\frac{1}{2}}$ for the Euler equations (ref. 8). Its components are computed as follows:

$$\rho_{j+\frac{1}{2}} = \frac{\rho_{j+1}\sqrt{\rho_j} + \rho_j\sqrt{\rho_{j+1}}}{\sqrt{\rho_j} + \sqrt{\rho_{j+1}}} \tag{8.5.2}$$

$$u_{j+\frac{1}{2}} = \frac{u_{j+1}\sqrt{\rho_j} + u_j\sqrt{\rho_{j+1}}}{\sqrt{\rho_j} + \sqrt{\rho_{j+1}}} \tag{8.5.3}$$

$$h_{j+\frac{1}{2}} = \frac{h_{j+1}\sqrt{\rho_j} + h_j\sqrt{\rho_{j+1}}}{\sqrt{\rho_j} + \sqrt{\rho_{j+1}}} \tag{8.5.4}$$

where,

$$h \equiv \frac{\gamma e}{\rho} - \frac{(\gamma - 1)(u_{j+\frac{1}{2}})^2}{2} \tag{8.5.5}$$

Naturally $\rho u_{j+\frac{1}{2}} \equiv \rho_{j+\frac{1}{2}} u_{j+\frac{1}{2}}$, and the final component, $e_{j+\frac{1}{2}}$ is computed using (8.5.5).

The discovery of this identity allows the construction of a TVD scheme for systems in an elegant way. First, the scheme is written as a central difference scheme with dissipation to be added later.

$$q_j^{n+1} = q_j^n - \frac{\Delta t}{\Delta x}(f_{j+1}^n - f_{j-1}^n) + \text{dissipation} \tag{8.5.6}$$

Next the flux difference is replaced with two differences

$$q_j^{n+1} = q_j^n - \frac{\Delta t}{\Delta x}\left[(f_{j+1}^n - f_j^n) + (f_j^n - f_{j-1}^n)\right] + \text{dissipation} \tag{8.5.7}$$

Roe's identity is used to replace the flux differences.

$$q_j^{n+1} = q_j^n - \frac{\Delta t}{\Delta x}\left[(f_{,q})_{j+\frac{1}{2}}(q_{j+1}^n - q_j^n) + (f_{,q})_{j-\frac{1}{2}}(q_j^n - q_{j-1}^n)\right] + \text{dissipation} \tag{8.5.8}$$

The quantities $(f_{,q})_{j+\frac{1}{2}}$ and $(f_{,q})_{j-\frac{1}{2}}$ are matrices in this case. They can be replaced by the usual eigensystem decomposition

$$\begin{aligned} q_j^{n+1} = q_j^n - \frac{\Delta t}{\Delta x}&(Y_{j+\frac{1}{2}}\Lambda_{j+\frac{1}{2}}Y^{-1}{}_{j+\frac{1}{2}}(q_{j+1}^n - q_j^n) \\ &+ Y_{j-\frac{1}{2}}\Lambda_{j-\frac{1}{2}}Y^{-1}{}_{j-\frac{1}{2}}(q_j^n - q_{j-1}^n) + \text{dissipation} \end{aligned} \tag{8.5.9}$$

At this point, if all the eigensystems were the same, (i.e., $Y_{j+\frac{1}{2}} = Y_{j-\frac{1}{2}} = Y_j$) then the equation would formally diagonalize into three nonlinear, scalar wave equations which could be treated just like Burgers' equation. In fact, this is not the case. This leads to an arbitrary choice in the pseudodiagonalization of the equations. Roe's choice is to treat the quantity ${Y^{-1}}_{j+\frac{1}{2}}(q^n_{j+1} - q^n_j)$ as a difference in a new variable $\tilde{q}$. The equivalent equation is

$$(q_{,t})^{n+1}_j = (q_{,t})^n_j - \frac{\Delta t}{\Delta x}(\Lambda_{j+\frac{1}{2}}(\tilde{q}_{j+\frac{1}{2}} - \tilde{q}_j) + \Lambda_{j-\frac{1}{2}}(\tilde{q}_j - \tilde{q}_{j-\frac{1}{2}})) \tag{8.5.10}$$

The TVD techniques described in Sections 8.3 and 8.4 can then be used to find $\tilde{q}_{j+\frac{1}{2}}$ and $\tilde{q}_{j-\frac{1}{2}}$. These are then multiplied by $Y_{j+\frac{1}{2}}$ and $Y_{j-\frac{1}{2}}$ respectively to recover $f^n_{j+\frac{1}{2}}$ and $f^n_{j-\frac{1}{2}}$. Finally, q is advanced in time using

$$q^{n+1}_j = q^n_j - \frac{\Delta t}{\Delta x}(f^n_{j+\frac{1}{2}} - f^n_{j-\frac{1}{2}}) \tag{8.5.11}$$

Other, similar schemes are possible. All require ignoring the difference between eigenvectors at nearby points. While numerical schemes for Burgers' equation allowed some ambiguity in the speed of convection, schemes for the Euler equations aren't even clear on precisely what is being convected.

The fundamental problem goes deeper than approximation inaccuracies. The TVD property doesn't hold for systems because of the interactions between the different eigenvectors. This is true for the PDE as well as for the numerical schemes discussed here.

8.6 Summary of the TVD Approach

There are some similarities between the TVD approach and the entropy approach of the first seven chapters. For the linear wave equation, Section 8.2 showed that certain TVD schemes also satisfy a cell entropy inequality. The PDE in fact possesses both properties. On the other hand, in Section 5.5 it was shown that at least one TVD scheme does not satisfy even a global entropy inequality, leading to unsatisfactory results. Overall it would appear that the second law of thermodynamics is the fundamental property and that the TVD property is simply an important side effect.

For Burgers' equation, Sections 8.3 and 8.4 showed that it is possible to construct schemes which rigorously satisfy the TVD conditions. In Section 6.3, a scheme was constructed which rigorously satisfied the cell entropy inequality, though this requires iteration at each cell and is very expensive. It is not clear whether such a scheme is rigorously TVD. It is clear from figure 6.7 that the particular TVD scheme used in Section 6.5 does not always satisfy a cell entropy inequality. This results in a shock that is slightly too strong at certain times, but the solution is otherwise quite satisfactory. Without the exact solution to compare with, the error would not have been detected. The main ambiguity in

nonlinear scalar equations of this type is the computation of a local wave speed and the variation of this wave speed over the cell. Indeed, it was shown that the assumption of a cellwise constant wave speed leads to the entropy errors just described.

For one-dimensional gasdynamics, it is not possible to construct consistent schemes which formally diminish total variation. This is because the physical problem does not have the TVD property. Many schemes have been constructed using TVD ideas by ignoring the interaction between the eigenvectors. These schemes appear to give excellent results, especially for reasonable ($M < 5$) Mach numbers, in spite of the approximations in their derivation. For hypersonic mach numbers ($M > 5$) the approximation of locally constant eigenstructure can become so bad that serious accuracy and convergence problems develop[1].

Although this section only deals with explicit TVD schemes, there has been considerable work done on implicit TVD schemes by Yee, Warming, and Harten (ref. 35). Such schemes produce results which are independent of the time step. Although the linearization is only an approximation to the gradient, it does allow larger time steps than the equivalent explicit method.

A further modification avoids the need to explicitly calculate the eigenvectors at each grid point. This is a considerable savings in computational work at the expense of slightly more dissipation. Known as symmetric TVD, it was originally proposed by Davis (ref. 36), and developed further by Yee(ref. 37), and others.

In two or three dimensions, the concept of total variation is not uniquely defined. Under one definition, the L_1 norm of the solution should diminish. It has been shown (ref. 38) that schemes which diminish this norm in more than one dimension cannot be more than first-order accurate in space. The usual fix is to split dimensionally, applying a one-dimensional TVD operator along each of the coordinate directions. Such schemes are second-order accurate, but are not formally TVD in multidimensions, even for scalars. Furthermore, they are not easily applied to unstructured, finite element style grids.

In problems with viscosity, heat transfer, or chemistry, there is no formal eigenvector decomposition as there is for the Euler equations. This makes it difficult to diagonalize the equations (even approximately) and apply the TVD property. The usual attack is to treat the inviscid convection terms separately and apply the TVD operator to them. The other terms are differenced in a symmetric stencil of suitable accuracy. The argument goes that such terms are inherently dissipative and hence can be relied upon not to create wiggles.

The problem of extending the cell entropy inequality ideas to multidimensions, unstructured grids, and more complete equation sets will be addressed in Chapter 10. There it will be shown that the second law extends nicely to all these cases.

[1] T. Barth, Some Notes on Shock Resolving Flux Functions, 1988 (unpublished)

Chapter 9. VARIOUS OTHER SCHEMES

9.1 Flux-Based Schemes

With the exception of the previous chapter, all schemes discussed so far have been variable-based schemes; i.e., schemes in which the numerical flux is determined by first computing the state of the independent variables at the midpoint. That is

$$f_{j+\frac{1}{2}} = f(q_{j+\frac{1}{2}}) \tag{9.1.1}$$

This approach allows the computation of an entropy flux, $F_{j+\frac{1}{2}}$ that is consistent with the conservative flux, $f_{j+\frac{1}{2}}$. With the exception of Godunov's scheme and its variants, there are no well-known schemes which use a flux of this type.

It is more common to construct a numerical flux as a combination of the fluxes computed at the grid points. That is

$$f_{j+\frac{1}{2}} = h(f_{j+1}, f_j) \tag{9.1.2}$$

A typical example is $f_{j+\frac{1}{2}} = \frac{1}{2}(f_j + f_{j+1})$, the central difference flux. Such schemes are referred to here as flux-based schemes.

Flux-based schemes are difficult to analyze from the point of view of a cell entropy inequality. This is because the entropy flux F is not a function of f. At a steady shock, for example, the conservative fluxes are constant through the shock, while the entropy flux jumps discontinuously.

Because F is not a function of f, it is difficult to compute $F_{j+\frac{1}{2}}$, which is needed to compute the entropy production rate for the cell. If entropy is considered at all, the usual approach is to independently approximate F. Although the global entropy production rate is unaffected by such an approximation, individual cells may have considerable errors. Such an approach is discussed in Section 9.3.

There is another way to analyze the entropy production of flux-based schemes, a way which involves changing the quadrature rules by which $f_{j+\frac{1}{2}}$ is computed. Recall from Chapter 3 that the underlying integral equations for a convection problem are

$$\int_V q_{,t}\, dv + \oint_{\partial V} f \cdot \mathbf{n}\, dA = 0 \tag{9.1.3}$$

Up to this point, one-dimensional problems have been treated as though the state were piecewise constant in each cell. In order to treat q as a continuous function, a profile has

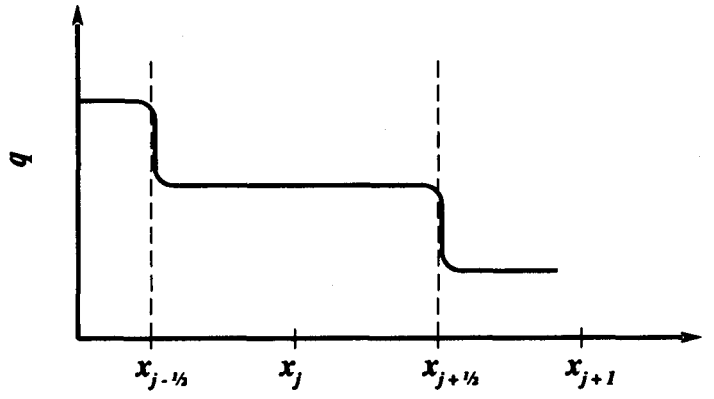

Figure 9.1. – The volume integral behaves as though $q(x)$ is piecewise constant over each control volume. Yet, the surface integrals are well defined.

been introduced between cells (see fig. 9.1). This profile is assumed to be sufficiently narrow that the volume integral is unaffected, that is

$$\int_V q_{,t}\, dv = (q_{,t})_j \Delta x \tag{9.1.4}$$

The surface integral however, is affected. The details of the profile determine the value of $q_{j+\frac{1}{2}}$, the state on the cell boundary. The surface integral is then evaluated as

$$\oint_{\partial V} f \cdot \mathbf{n}\, dA = f(q_{j+\frac{1}{2}}) - f(q_{j-\frac{1}{2}}) \tag{9.1.5}$$

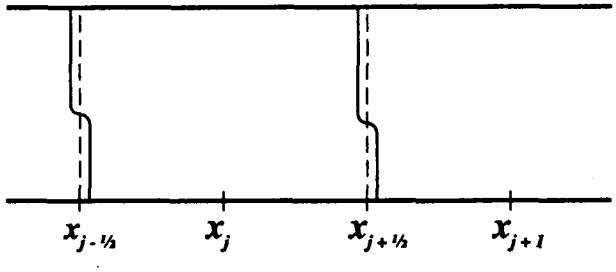

Figure 9.2. – In this view of a quasi-one-dimensional problem, $q(x, y)$ is piecewise constant over a slightly different area than the control volume. The thin, solid, curved lines are contours of q. The surface integrals (carried out on the dotted lines) must be evaluated in two parts.

It is often helpful to view one-dimensional processes as a limit of two-dimensional processes. For example, consider the convection in a two-dimensional pipe of unit cross section (fig. 9.2). If $q = q(x)$, $q_{,y} = 0$, then a one-dimensional flow is described. Now, instead of allowing a profile to connect the states between cells, suppose the requirement that $q_{,y} = 0$ is relaxed. Let the state be constant along a slightly different coordinate line, η, i.e., $q_{,\eta} = 0$. This is shown as the thin solid line in figure 9.2. Assume that this line is sufficiently close to the cell boundary that the volume integral remains unaffected and can be evaluated as before.

The surface integral contains one term for each cell face, as before.

$$\oint_{\partial V} f \cdot \mathbf{n}\, dA = f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}} \tag{9.1.6}$$

In 9.1.5, $f_{j+\frac{1}{2}}$ was evaluated easily because $q_{j+\frac{1}{2}}$ was a constant along the cell face. The situation shown in figure 9.2 is different, requiring the integration along each cell face to be done in two parts.

$$\begin{aligned} f_{j+\frac{1}{2}} &= \frac{1}{2}\left[(2-\phi_{j+\frac{1}{2}})f(q_j) + \phi_{j+\frac{1}{2}} f(q_{j+1})\right] \\ &= \frac{1}{2}\left[(2-\phi_{j+\frac{1}{2}})f_j + \phi_{j+\frac{1}{2}} f_{j+1}\right] \end{aligned} \tag{9.1.7}$$

The parameter $\phi_{j+\frac{1}{2}}$ depends on the geometrical details shown in figure 9.2, details which are adjustable. The flux $f_{j+\frac{1}{2}}$ varies linearly between f_j and f_{j+1} as $\phi_{j+\frac{1}{2}}$ goes from 0 to 2. When $\phi_{j+\frac{1}{2}} = 1$, central differencing is recovered.

The advantage of the physical point of view shown in figure 9.2 is that the entropy flux can now be computed. It is integrated along the same path as the other fluxes, resulting in the expression

$$F_{j+\frac{1}{2}} = \frac{1}{2}\left[(2-\phi_{j+\frac{1}{2}})F_j + \phi_{j+\frac{1}{2}} F_{j+1}\right] \tag{9.1.8}$$

where $F_{j+1} = F(q_{j+1})$, etc. This allows the cell entropy production rate to be computed. It is sometimes possible to select $\phi_{j+\frac{1}{2}}$ in such a way that this $(\dot{P}_s)_j \geq 0$.

For an example, consider the wave equation again. What is the entropy production rate, given this new definition of F? (Since $f = cu$ in this case, the new definition of f has no effect.) Following the same derivation path as before (Section 5.1), the semi-discrete entropy production rate is

$$\Delta x(\dot{P}_s)_j = -c\left[\phi_{j+\frac{1}{2}}(u_{j+1}-u_j)^2 - (2-\phi_{j-\frac{1}{2}})(u_j-u_{j-1})^2\right] \geq 0 \tag{9.1.9}$$

As before, the substitution $r_j = \frac{u_j - u_{j-1}}{u_{j+1}-u_j}$ is helpful. For positive wave speeds this leads to

$$\frac{\phi_{j+\frac{1}{2}}}{{r_j}^2} \leq 2 - \phi_{j-\frac{1}{2}} \tag{9.1.10}$$

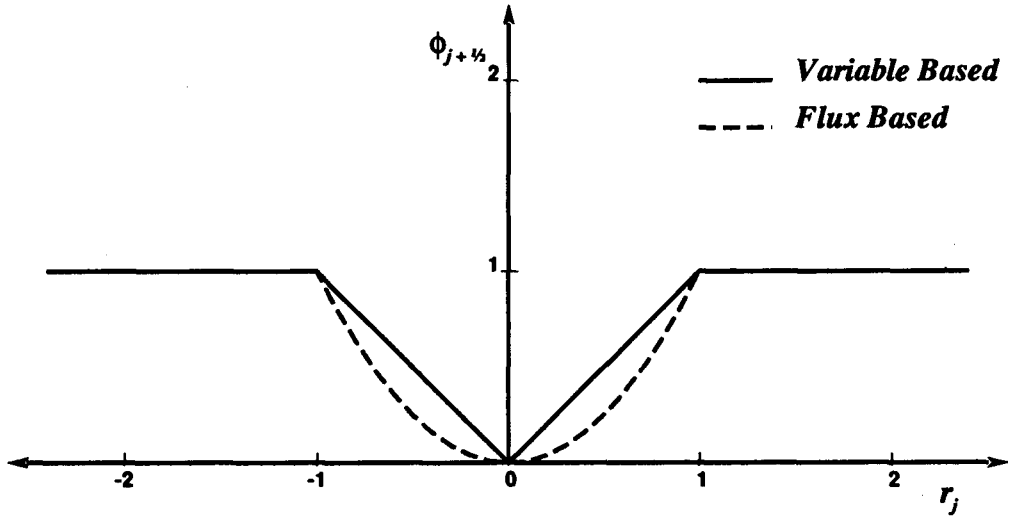

Figure 9.3. – Both limiters are second-order accurate. Both satisfy an entropy inequality. The variable-based limiter has about half the dissipation of the flux-based limiter.

which is satisfied if

$$\phi_{j+\frac{1}{2}} \equiv \min({r_j}^2, 1) \tag{9.1.11}$$

For comparison, the equivalent expression for the variable-based scheme is

$$\phi_{j+\frac{1}{2}} \equiv \min(|r_j|, 1) \tag{9.1.12}$$

These two limiters are plotted in figure 9.3. Both limiters are asymptotically second-order accurate because both are continuous through the point (1,1) on the graph. However, the flux-based limiter is about twice as dissipative as the variable-based limiter.

For the wave equation with $c = 1$, $f_{j+\frac{1}{2}} = u_{j+\frac{1}{2}}$. Thus, every flux-based scheme can be viewed as a variable-based scheme and *vice versa*. Both limiters are stable and give satisfactory results. Thus, the difference between the two is purely one of analysis.

Another problem arises when flux-based schemes are extended to systems of equations. The view shown in figure 9.2 allows only one interpolation parameter at each cell. This is too restrictive to guarantee positive entropy production rates at sonic points. Numerical experiments revealed significant problems at such points in the form of oscillations, glitches and instabilities.

Flux-based schemes, as defined in this section, are an interesting dead-end. To make any further progress, it is necessary to somehow remove the restriction of having only one interpolation parameter per cell. While this is possible conceptually, rigorous analysis is difficult.

9.2 Flux Splitting and the Second Law

An interesting connection exists between flux splitting and the second law of thermodynamics. It involves a Taylor series approximation and is therefore inexact. Nevertheless, some significant insights are possible.

Formally, flux splitting only applies to certain purely hyperbolic systems of equations such as the Euler equations. These are written in the usual way

$$q_{,t} + f_{,x} = 0 \tag{9.2.1}$$

The semi-discrete equations can be written

$$q_{,t} + \frac{1}{\Delta x}(f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}) = 0 \tag{9.2.2}$$

The general idea behind flux splitting is to use the homogeneous property ($f_j = (f_{,q})_j q_j$) or Roe's identity (8.5.1), and the known eigensystem of $f_{,q}$ (7.3.2), to approximate a system of equations by several scalar equations. Methods which use $f_j = (f_{,q})_j q_j$ are known as flux vector splitting methods while those which use Roe's identity are known as flux difference splitting methods.

As previously derived (3.4.3), the semi-discrete cell entropy inequality is equivalent to

$$[-(S_{,q})_j(f_{j+\frac{1}{2}} - f_j) + (F_{j+\frac{1}{2}} - F_j)] + [-(S_{,q})_j(f_j - f_{j-\frac{1}{2}}) + (F_j - F_{j-\frac{1}{2}})] \geq 0 \tag{9.2.3}$$

The quantity on the left is the entropy production rate for the j^{th} cell. In smooth regions it should approximately vanish. This notion can be formalized by expanding in a Taylor series about q_j. For example,

$$f_{j+\frac{1}{2}} \approx f_j + (f_{,q})_j(q_{j+\frac{1}{2}} - q_j) + \frac{1}{2}(q_{j+\frac{1}{2}} - q_j)^T(f_{,qq})_j(q_{j+\frac{1}{2}} - q_j) \tag{9.2.4}$$

The term $(q_{j+\frac{1}{2}} - q_j)$ appears so often in expansions of this type that it is replaced with the letter a. Similarly, $b \equiv (q_j - q_{j-\frac{1}{2}})$. Finally, the subscript j, now redundant, is dropped. Grouping the terms by powers of a and b, (9.2.3) becomes

$$\begin{aligned}
&[-S_{,q}(f_j - f_j) + (F_j - F_j)] \\
&-[-S_{,q}(f_j - f_j) + (F_j - F_j)] \\
&+[-S_{,q}f_{,q}(a) + F_{,q}(a)] \\
&-[-S_{,q}f_{,q}(b) + F_{,q}(b)] \\
&+\frac{1}{2}[-S_{,q}(a^T f_{,qq}a) + a^T F_{,qq}a] \\
&-\frac{1}{2}[-S_{,q}(b^T f_{,qq}b) + b^T F_{,qq}b] \geq 0
\end{aligned} \tag{9.2.5}$$

The first two terms vanish by inspection. The second two terms vanish because of the compatability condition, $F_{,q} = S_{,q} f_{,q}$. Cancelling the factor of $\frac{1}{2}$, (9.2.5) becomes

$$[a^T(F_{,qq} - S_{,q} f_{,qq})a] - [b^T(F_{,qq} - S_{,q} f_{,qq})b] \geq 0 \tag{9.2.6}$$

This simplifies substantially since

$$F_{,qq} = (F_{,q})_{,q} = (S_{,q} f_{,q})_{,q} = S_{,qq} f_{,q} + S_{,q} f_{,qq} \tag{9.2.7}$$

which comes from the compatability condition and the chain rule. This cancels the $f_{,qq}$ term (a third-rank tensor) and (9.2.6) collapses to

$$a^T S_{,qq} f_{,q} a - b^T S_{,qq} f_{,q} b \geq 0 \tag{9.2.8}$$

Since $S_{,qq}$ is negative definite, it is clear that $f_{,q}$ plays a key role in the satisfaction of the second law. This observation is clarified if the identity (7.3.10) is applied. Recall, this is

$$S_{,qq} f_{,q} = -(Y^{-1})^T \Lambda Y^{-1} \tag{9.2.9}$$

Inequality (9.2.6) can be formally diagonalized using this identity. This allows the two terms to be combined and leads to the expression

$$\begin{aligned} &\lambda_1(\tilde{a}_1^2 - \tilde{b}_1^2) \\ &+\lambda_2(\tilde{a}_2^2 - \tilde{b}_2^2) \\ &+\lambda_3(\tilde{a}_3^2 - \tilde{b}_3^2) \geq 0 \end{aligned} \tag{9.2.10}$$

Satisfaction of (9.2.10) is a fairly simple matter, accomplished on a term-by-term basis. Thus, a system of equations is broken into three, nonlinear, scalar equations. The resulting scheme is identical to flux vector splitting (ref. 9). This allows the analysis of flux vector splitting from the point of view of the second law.

In Chapter 6, the behavior of nonlinear scalar equations was studied. In particular, figure 6.2 show that the approximations $u_{j+\frac{1}{2}} = u_j$ and $u_{j+\frac{1}{2}} = u_{j+1}$ may both violate a cell entropy inequality at sonic points. First-order flux vector splitting is constrained to use one of these approximations. It seems likely that a local violation of the second law is responsible for the sonic point glitch which is characteristic of first-order flux vector splitting.

As a second-order scheme, $u_{j+\frac{1}{2}}$ will approach the average of u_j and u_{j+1} as the mesh is refined. Since the Taylor series is taken about u_j, the Taylor Series errors do not vanish. Thus, the validity of this analysis for second-order flux vector splitting is questionable.

Flux difference splitting, on the other hand is generally linearized about a state that is, in some sense, the average of u_j and u_{j+1} (perhaps the Roe average of Section 5.8). In addition to avoiding the sonic point problems experienced by flux vector splitting, the linearization error vanishes rapidly in the limit of mesh refinement.

It is worth noting here that attempts were made (unsuccessfully) to construct a scheme which satisfied a cell entropy inequality using (9.2.10). Because of truncation errors, (9.2.10) does not imply satisfaction of (9.2.3). These discrepancies proved large enough to cause instabilities.

9.3 Tadmor's Entropy Scheme

So far, the emphasis of this work has been on the satisfaction of the second law in each cell. On the other hand the only formal theorem involving entropy requires satisfying the second law over the whole domain. This section examines a scheme which satisfies a global entropy inequality without satisfying a local entropy inequality.

Tadmor showed (ref. 15) that many equations of interest can be discretized in a way that globally conserves entropy, in addition to mass momentum and energy. He does not recommend such schemes but shows how to construct them. He also gives a fairly general form of dissipation which results in positive (global) entropy production rates. Such schemes fall into a class he terms entropy-stable schemes.

In this section, the limiting case of entropy conservation is explored. Since Burgers' equation produces entropy at shocks, an entropy conserving scheme is certain to be at variance with the physics. It is interesting to observe the behavior of such a scheme.

The equation of interest here is Burgers' equation.

$$u_{,t} + \frac{1}{2}(u^2)_{,x} = 0 \tag{9.3.1}$$

By carefully following the analysis in (ref. 15), the following flux is derived

$$f_{j+\frac{1}{2}} = \frac{1}{6}({u_j}^2 + u_j u_{j+1} + {u_{j+1}}^2) \tag{9.3.2}$$

It is proved by Tadmor that this flux conserves entropy in the sense that

$$(S_{,t})_j + \frac{1}{\Delta x}[{F^*}_{j+\frac{1}{2}} - {F^*}_{j-\frac{1}{2}}] = 0 \tag{9.3.3}$$

The quantity ${F^*}_{j+\frac{1}{2}}$ (given in (ref. 15)) is an approximation to the entropy flux. It is given in general by the expression

$${F^*}_{j+\frac{1}{2}} = \frac{1}{2}[F_{j+1} - (S_{,q})_{j+1}(f_{j+1} - f_{j+\frac{1}{2}})] + \frac{1}{2}[F_j + (S_{,q})_j(f_{j+\frac{1}{2}} - f_j)] \tag{9.3.4}$$

which for Burgers' equation reduces to

$$F^*_{j+\frac{1}{2}} = -\frac{1}{3}u_j u_{j+1}(u_{j+1} + u_j) \tag{9.3.5}$$

Notice that this becomes exact, as does $f_{j+\frac{1}{2}}$ when $u_{j+1} = u_j$. Thus, both $F^*_{j+\frac{1}{2}}$ and $f_{j+\frac{1}{2}}$ meet the Lax condition for consistency. Nevertheless they are independently approximated. That is, there is no $u_{j+\frac{1}{2}}$ for which, simultaneously, $f_{j+\frac{1}{2}} = f(u_{j+\frac{1}{2}})$ and $F^*_{j+\frac{1}{2}} = F(u_{j+\frac{1}{2}})$.

Using implicit Euler time advance to advance the semi-discrete scheme of (9.3.2) ensures that the fully discrete entropy production rate will be greater than the semi-discrete entropy production rate (see Section 4.3). The results for this scheme are shown in figures 9.4 and 9.5. Figure 9.4 shows the excellent agreement between Tadmor's scheme and the analytical solution prior to the formation of a shock. On the other hand, the post-shock performance is terrible. Figure 9.5 shows the computed versus exact solution at the time of peak shock strength. Notice the severe oscillations in the neighborhood of the shock.

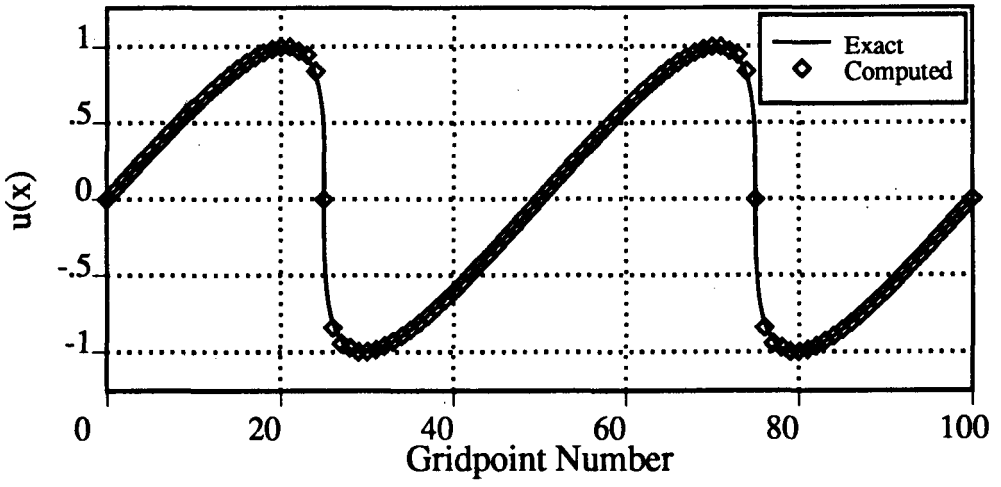

Figure 9.4. – Tadmor's solution at onset of shock.

In spite of the oscillations and generally poor quality, the solution remains bounded. In particular, the sum of all the u_j's is conserved as is the sum of all the ${u_j}^2$'s. Since $S \equiv -u^2$ for this problem the second law is satisfied globally. The question of cell-by-cell entropy production remains.

Although (9.3.4) is a good approximation to the entropy flux in smooth regions, it is a poor approximation near the shock. It is also a poor approximation amongst the wiggles.

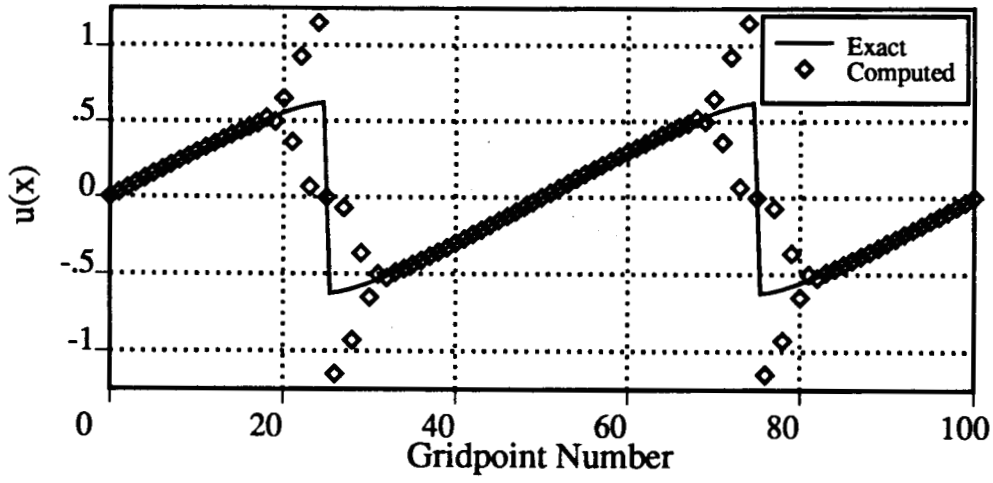

Figure 9.5. – Tadmor's solution at maximum shock strength. Both u and u^2 are conserved quantities when, in fact, u^2 should be decreasing. Hence the wiggles.

To show this, one may compute an exact $u_{j+\frac{1}{2}}$ by computing $f_{j+\frac{1}{2}}$ from (9.3.2) and then solving $f_{j+\frac{1}{2}} = \frac{1}{2}{u_{j+\frac{1}{2}}}^2$ for $u_{j+\frac{1}{2}}$. The sign is given by consistency, that is $u_{j+\frac{1}{2}}$ has the same sign as $u_j + u_{j+1}$. Once the exact $u_{j+\frac{1}{2}}$ is known, $F_{j+\frac{1}{2}}$ can be computed using the formula

$$F_{j+\frac{1}{2}} = -\frac{2}{3}{u_{j+\frac{1}{2}}}^3 \tag{9.3.7}$$

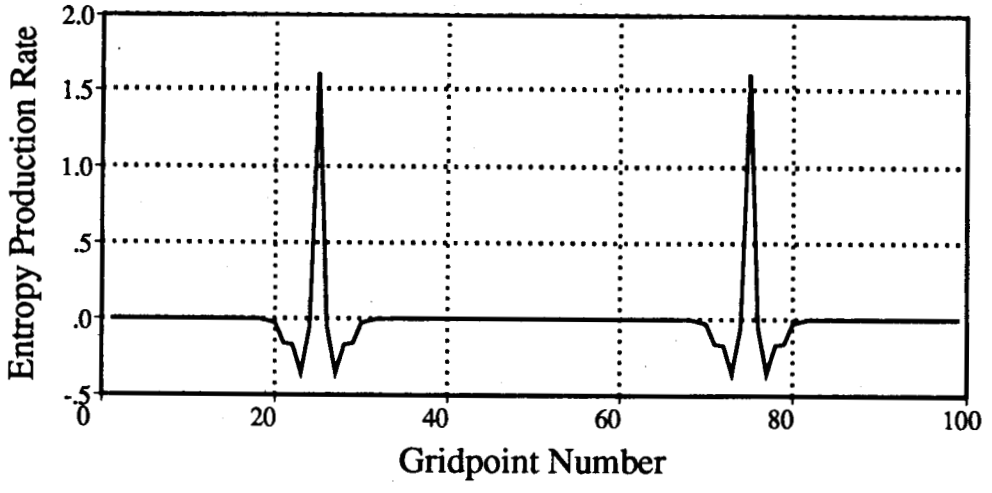

Figure 9.6. – Entropy production rate for figure 9.5. Notice that no net entropy is produced over the domain. There are individual cells that destroy entropy. These are the source of the wiggles in u.

With this definition for $F_{j+\frac{1}{2}}$, it is possible to evaluate the true entropy production rate, cell-by-cell, as in Chapter 6. This is shown in figure 9.6 for the solution shown in figure 9.5. The total entropy production rate for the domain is zero, but the entropy production rate in some cells is negative. The entropy production rate computed by Tadmor is positive and very small, invisible on this scale, all of it due to the time advance.

This discrepancy between two ways of computing entropy production rates is only evident in the presence of shocks. It is significant that (9.3.2) works quite well until a shock appears. At the onset of the shock, two things happen at once; oscillations appear in the solution and cells near the shock begin to display a negative entropy production rate.

It must be noted that neither Tadmor or Osher ever intended for the scheme to be used in this way. In particular, Osher recommends the addition of a suitable limiter, to enforce a TVD condition. Along the same lines, Tadmor suggests the addition of a dissipative term of a particular form. This form is sufficiently general to allow the limiter recommended by Osher. Either way, the added dissipation increases the global entropy production rate. This (according to Tadmor (ref. 15)) allows the scheme to converge to the correct answer as the mesh is refined.

The scheme is nevertheless instructive as written. It shows the one-to-one correspondence between local violations of the second law and unphysical features of the flow (oscillations). This in turn shows the importance of satisfying a local entropy condition in addition to a global one.

9.4 Entropy and Wiggles

The nonlinear stability conjecture of Chapter 2 depends on satisfaction of the second law, but only in a global sense. Throughout this work, the much more restrictive cell entropy inequality has been stressed. Certainly, if entropy is produced in each cell, it is produced in the domain as a whole. Thus, the satisfaction of a cell entropy inequality implies nonlinear stability, at least for scalars.

It is definitely easier to construct schemes which only have to satisfy a global entropy inequality. For example it should be fairly simple to solve for $q_{j+\frac{1}{2}}$ such that $(\dot{P}_s^+)_j = -(\dot{P}_s^-)_{j+1}$, since both of these depend only on $q_{j+\frac{1}{2}}$. Yet, it only makes sense to enforce the first and second law on the same control volumes, the computational cells.

Perhaps the cell entropy inequality provides some desirable property in addition to stability, in exchange for the increased computational complexity. There are indications that this might be so. Recall that in Chapter 5, the schemes which satisfied a cell entropy inequality turned out to be TVD as well. The previous section showed a scheme which did not satisfy a cell entropy inequality but did satisfy the second law in a global sense.

The result was a scheme which was stable as promised, but with severe oscillations in the solution. This suggests that the cell entropy inequality provides control of oscillations, in addition to stability.

Consider now a further illustration of this hypothesis. The problem considered was previously discussed in Section 7.6. The solution, given in figure 9.7, was obtained with a flux-based central difference scheme using a constant coefficient, second difference, dissipation model. The analytical solution for the density is given by the solid line while the computed solution is plotted as symbols. Notice the overshoots in the neighborhood of the shock. Intuitively these seem to be unphysical, yet the method is fully conservative of mass, momentum, and energy. Unfortunately, it does not satisfy the second law.

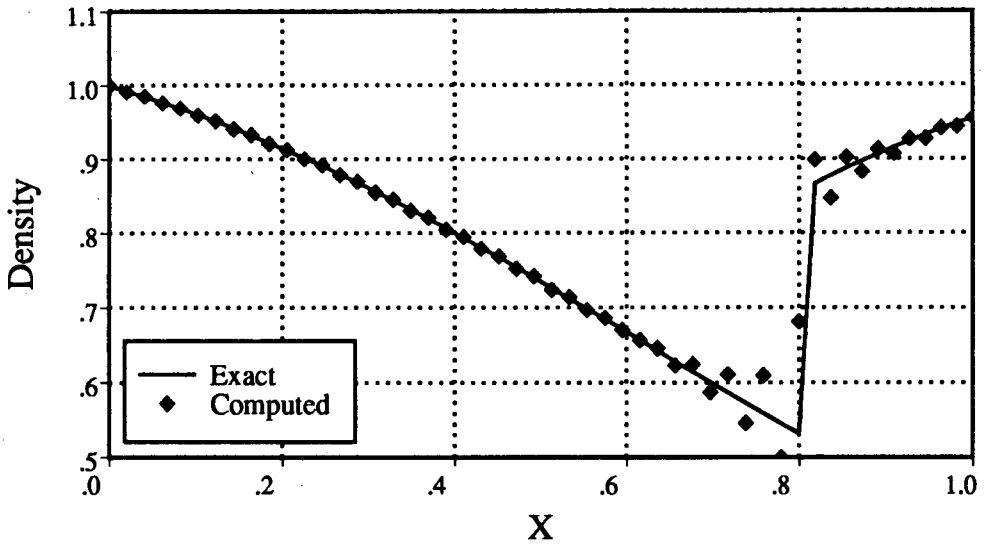

Figure 9.7. – Compare this density distribution for a primitive one-dimensional Euler solution with that shown in figure 7.6. Notice the oscillations in the computed solution around the shock.

For steady solutions the entropy inequality simplifies to

$$\dot{P}_s = \oint_{\partial V} \mathbf{F} \cdot \mathbf{n}\, dA \geq 0 \tag{9.4.1}$$

where the quantity $\mathbf{F}$ is the local entropy flux. This allows an after-the-fact evaluation of solutions to the steady one-dimensional Euler equations with regard to their satisfaction of the second law. For this case the entropy flux F is $(\rho ua)s$ where s is the specific entropy of the working fluid and ρua is the mass flow rate. The second law in this case amounts to a requirement of monotonicity of s, since ρua is a constant for steady state solutions.

For steady solutions then, equation (9.4.1) should be obeyed over any control volume. The integrals reduce to differences in the one-dimensional case. By evaluating $\dot{P}_s$ for control volumes bounded by adjacent grid points, it is possible to see whether this method satisfies a cell entropy inequality. In figure 9.8, the entropy flux is given.

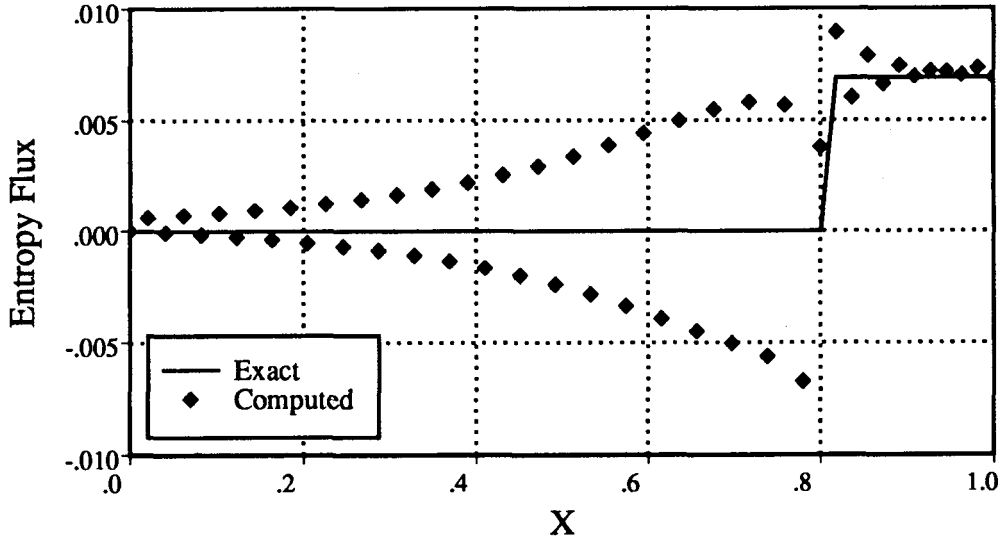

Figure 9.8. – Plot of entropy flux for a primitive Euler solution. The computed entropy flux is not a monotone increasing function of x. Therefore this solution does not satisfy the second law. Notice that the oscillations in entropy flux extend much farther than the oscillations in density.

As in figure 9.7, the symbols represent the computed solution. It is immediately apparent that the solution fails to satisfy the second law (F declines) in approximately half the cells. It is possible to group some of the control volumes together to form new control volumes, each of which obeys the second law. Some grid points fall entirely within these new control volumes. Such points are considered numerical artifacts and are not plotted. It is possible to choose a (nonunique) subset of the data which does satisfy the steady state second law for each cell (i.e., one for which F is monotone increasing). Figure 9.9 shows the entropy flux for such a subset.

In figure 9.10, the corresponding density values are plotted for these points. The wiggles are gone and the solution doesn't look too bad, though the shock is smeared.

In summary, this numerical scheme uses grid points inefficiently. The solution has roughly the same information content as a more dissipative scheme with half as many points.

Incidentally, the choice of which half of the solution to use is not arbitrary as might

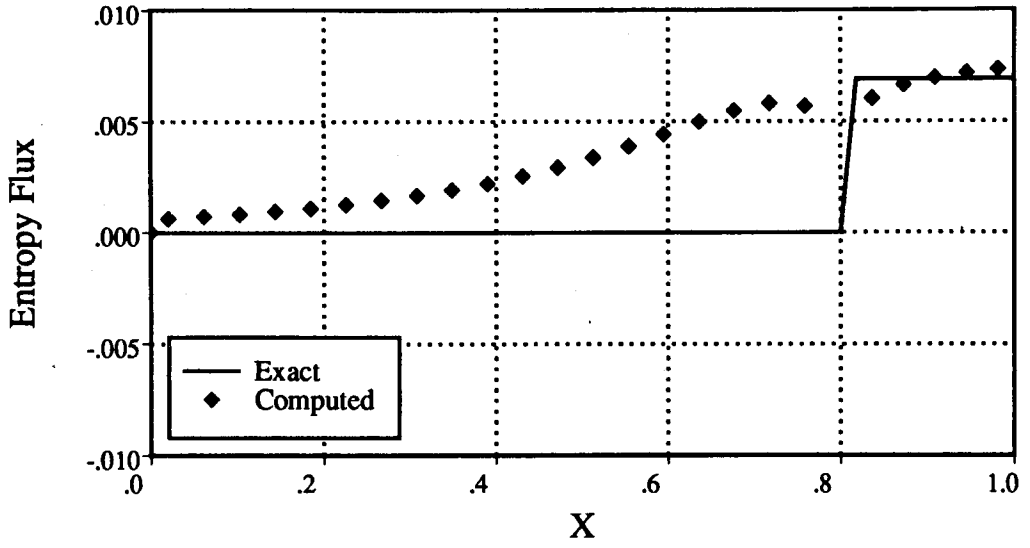

Figure 9.9. – Plot of entropy flux for selected grid points. These points form a set for which entropy flux is monotone increasing. The second law is satisfied over these points.

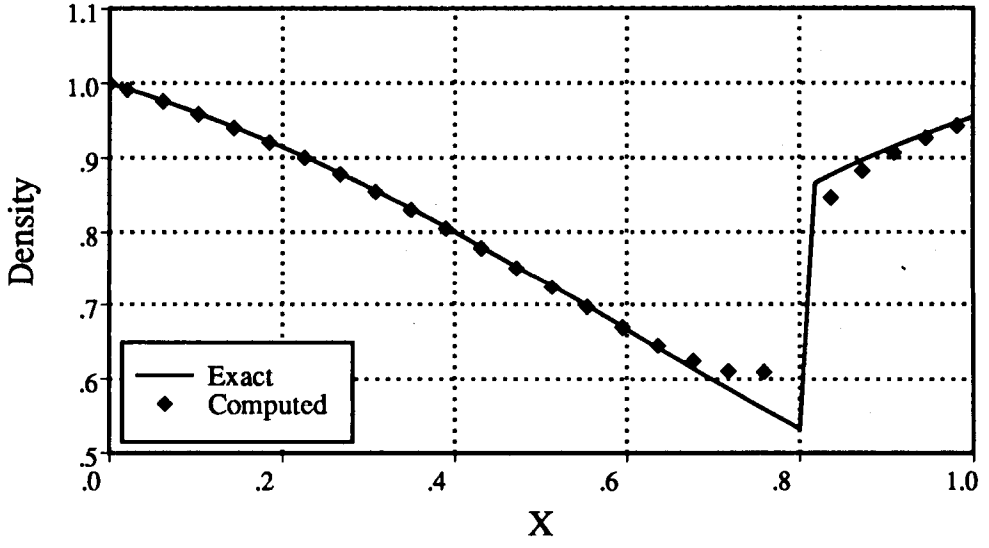

Figure 9.10. – Plot of density for selected grid points. This is the same set of points as in figure 9.9. Notice the absence of oscillations around the shock.

be suggested by an odd-even decoupling argument. The points which are excluded from figure 9.10 are shown in figure 9.11. Although the solution appears reasonable, and has a sharp shock, it represents a physically impossible flow. The compression requires less work than the theoretical minimum and the post-shock expansion produces more work

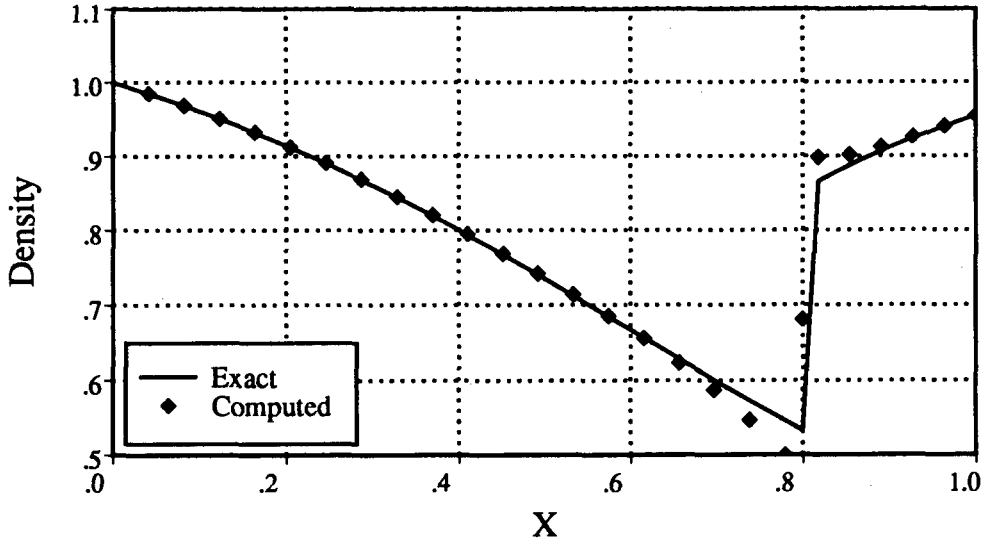

Figure 9.11. – The density values excluded from figure 9.10. These values form a smooth set. However the shock is too strong and the minimum density is too low. This could lead to negative densities in practice.

than the theoretical maximum. All this excess availability is consumed in the shock, which is noticeably stronger than it should be. As a practical matter, the minimum density is significantly low. When capturing a very strong shock, this could lead to negative densities or pressures. Very few schemes can tolerate this unphysical situation.

Increasing the dissipation coefficient sufficiently will give a solution without any wiggles on the original grid. Such a solution is shown in figure 9.12. When the entropy flux is plotted in figure 9.13, it becomes apparent that the entropy inequality is not satisfied over a large part of the domain.

The trick used in the previous example would require removal of the entire supersonic region (indeed the accuracy is quite poor in that region). Thus, smooth solutions do not necessarily represent the physics accurately. They may contain less information than solutions with oscillations.

At least for this example, the hypothesis appears to hold. If a solution satisfies, in each cell, conservation of mass momentum and energy and the second law, it will not exhibit unphysical oscillations. Such a principle is very useful in constructing numerical schemes.

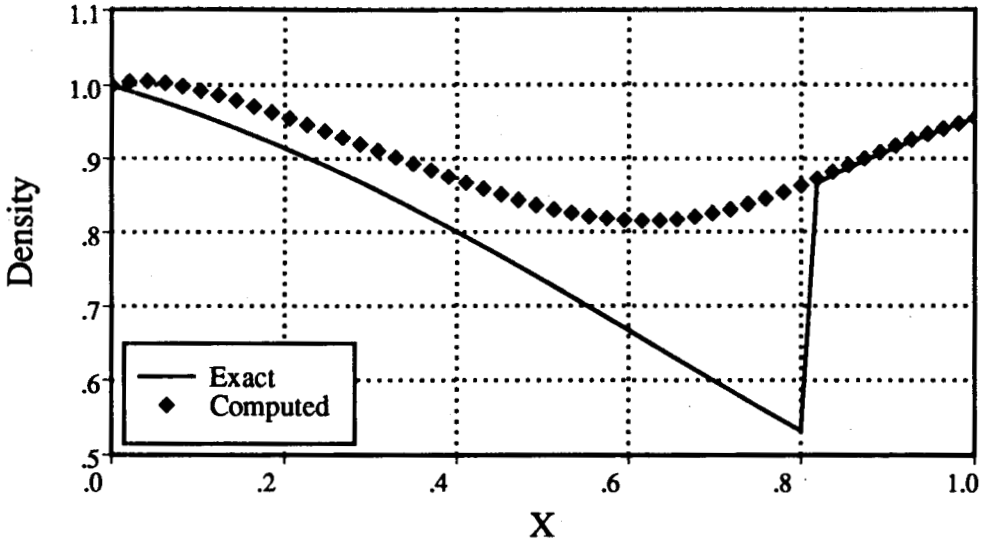

Figure 9.12. – Density distribution using a highly dissipative scheme. This solution is smooth, but highly inaccurate.

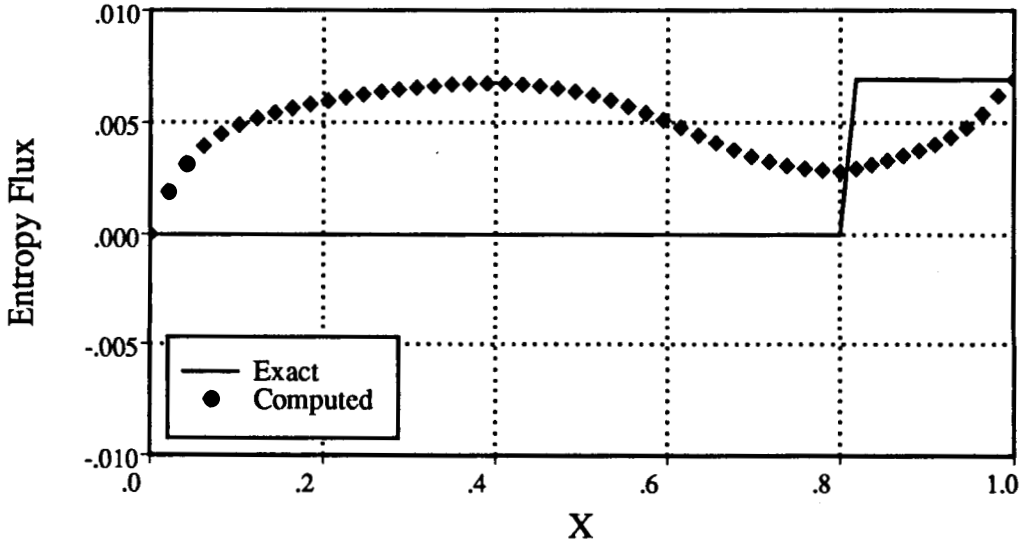

Figure 9.13. – Plot of entropy flux corresponding to figure 9.12. Although the solution in figure 9.12 is smooth, it does not satisfy the second law. Smoothness is not enough, even with conservation.

Chapter 10. PROPOSED EXTENSIONS AND SOME OBSERVATIONS

The discipline of thermodynamics is widely applicable. Essentially it applies over the whole of continuum mechanics. This includes phenomena involving heat transfer, viscosity, and chemistry, in two or three dimensions.

The numerical idea of a cell entropy inequality derives from the analytic ideas of thermodynamics. It seems reasonable that it should be just as widely applicable. This section discusses the details of how various extensions might be implemented.

10.1 Multidimensions

If the numerical techniques described in previous sections are to be useful, they must be extended to more than one dimension. This can, of course, be done in the usual way — by restricting the computation to regular grids and applying the one-dimensional algorithm in each dimension separately. This technique has widely recognized difficulties. Fortunately there is a more satisfying approach available.

Recall from Chapter 3 that the particular equations of interest, in integral form, are the conservation law

$$\int_V q[t+\Delta t]\,dv - \int_V q[t]\,dv + \int_t^{t+\Delta t} \oint_{\partial V} \mathbf{f}\cdot\mathbf{n}\,dA\,d\tau = 0 \qquad (10.1.1)$$

and the corresponding statement of the second law

$$\int_t^{t+\Delta t}\int_V \dot{P}_s\,dv\,d\tau \equiv \int_V S[t+\Delta t]\,dv - \int_V S[t]\,dv + \int_t^{t+\Delta t}\oint_{\partial V}\mathbf{F}\cdot\mathbf{n}\,dA\,d\tau \geq 0 \qquad (10.1.2)$$

As a first step, assume that the domain of interest is tessellated into a finite number of control volumes, each of which will therefore have a finite size. Assuming quadrilateral cells arranged in a regular array as in figure 10.1, these can be identified by the two indices j and k. The surface surrounding this control volume is also identified by the indices jk.

The integral equation which applies to this volume is

$$\int_{V_{jk}} q[t+\Delta t]\,dv - \int_{V_{jk}} q[t]\,dv + \int_t^{t+\Delta t}\oint_{\partial V_{jk}}\mathbf{f}\cdot\mathbf{n}\,dA\,d\tau = 0 \qquad (10.1.3)$$

Because each control volume has a finite extent, the volume integral must be a continuous function of time. Dividing (10.1.3) by Δt and proceeding to the limit $\Delta t \to 0$ gives the semi-discrete form.

$$\frac{\mathrm{d}}{\mathrm{dt}}\int_{V_{jk}} q\,dv + \oint_{\partial V_{jk}}\mathbf{f}\cdot\mathbf{n}\,dA = 0 \qquad (10.1.4)$$

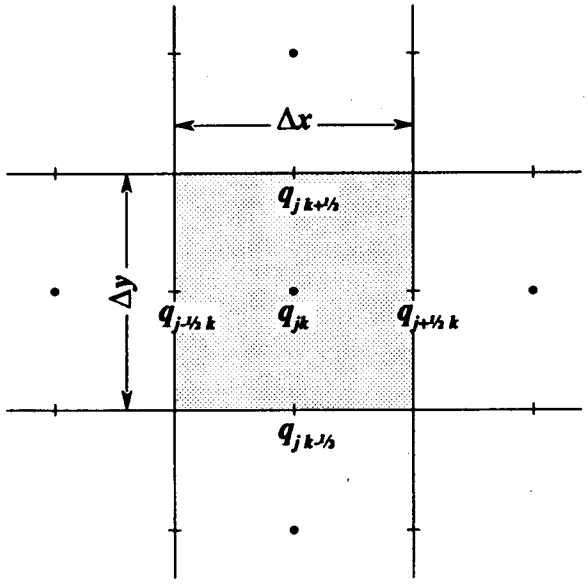

Figure 10.1. – A typical control volume in two dimensions.

Equation (10.1.4) holds at each time. At this point, the equations can be simplified by a definition.

$$q_{jk} \equiv \frac{1}{V_{jk}} \int_{V_{jk}} q \, dv = \overline{q}_{jk} \tag{10.1.5}$$

With this definition, (10.1.4) becomes

$$v_{jk} \frac{dq_{jk}}{dt} + \oint_{\partial V_{jk}} \mathbf{f} \cdot \mathbf{n} \, dA = 0 \tag{10.1.6}$$

In one dimension the surface integrals could be carried out exactly. This is not true in two dimensions. It is possible to approximate these integrals however. The control surface, ∂V_{jk}, which surrounds the control volume, can be broken into four segments in a straightforward way. As labeled in the figure, q is assumed to be constant along each of the four segments. Referring to figure 10.1, these four values are $q_{j+\frac{1}{2}\,k}$, $q_{j\,k+\frac{1}{2}}$, $q_{j-\frac{1}{2}\,k}$, and $q_{j\,k-\frac{1}{2}}$. Subject to this assumption

$$\begin{aligned} \oint_{\partial V_{jk}} \mathbf{f} \cdot \mathbf{n} \, dA = {} & a_{j+\frac{1}{2}\,k} \mathbf{f}_{j+\frac{1}{2}\,k} \cdot \mathbf{n} + a_{j-\frac{1}{2}\,k} \mathbf{f}_{j-\frac{1}{2}\,k} \cdot \mathbf{n} \\ & + a_{j\,k+\frac{1}{2}} \mathbf{f}_{j\,k+\frac{1}{2}} \cdot \mathbf{n} + a_{j\,k-\frac{1}{2}} \mathbf{f}_{j\,k-\frac{1}{2}} \cdot \mathbf{n} \end{aligned} \tag{10.1.7}$$

where in each term, $\mathbf{n}$ is an outward facing unit normal at the appropriate face. The cell face areas, e.g., $a_{j+\frac{1}{2}\,k}$, appear separately in each term, allowing for a more general grid than that shown. Combining (10.1.6) and (10.1.7), and dividing by the volume (also called

the metric Jacobian) gives

$$(q_{jk})_{,t} + \frac{a_{j+\frac{1}{2}\,k}}{v_{jk}}\mathbf{f}_{j+\frac{1}{2}\,k}\cdot\mathbf{n} + \frac{a_{j-\frac{1}{2}\,k}}{v_{jk}}\mathbf{f}_{j-\frac{1}{2}\,k}\cdot\mathbf{n} + \frac{a_{j\,k+\frac{1}{2}}}{v_{jk}}\mathbf{f}_{j\,k+\frac{1}{2}}\cdot\mathbf{n} + \frac{a_{j\,k-\frac{1}{2}}}{v_{jk}}\mathbf{f}_{j\,k-\frac{1}{2}}\cdot\mathbf{n} = 0 \tag{10.1.8}$$

For the rectangular mesh shown in figure 10.1, certain simplifications are possible. Using the traditional unit vectors $\mathbf{i}$ and $\mathbf{j}$ in the x and y directions, the dot product $\mathbf{f}\cdot\mathbf{n}$ can be carried out. Since the cell faces are aligned with coordinate directions, the following definitions are sufficient.

$$f \equiv \mathbf{f}\cdot\mathbf{i} \qquad \text{and} \qquad g \equiv \mathbf{f}\cdot\mathbf{j} \tag{10.1.9}$$

Since the various areas are Δx or Δy and the cell volume is $\Delta x \Delta y$, equation (10.1.8) can be written, for rectangular meshes, as

$$(q_{jk})_{,t} + \frac{f_{j+\frac{1}{2}\,k} - f_{j-\frac{1}{2}\,k}}{\Delta x} + \frac{g_{jk+\frac{1}{2}} - g_{jk-\frac{1}{2}}}{\Delta y} = 0 \tag{10.1.10}$$

This is analogous to equation (3.1.6), the one-dimensional version. The approximation in (10.1.7) remains, (10.1.10) is not exact. This is the major difference between one dimension and multidimensions.

The piecewise constant approximation used in one dimension is retained.

$$q = \overline{q}_{jk} \quad \text{for all } q \text{ within control volume } jk \tag{10.1.11}$$

where $\overline{q}_{jk}$ is the average value of q over the control volume. As shown in Section 2.2, this assumption maximizes the entropy within the control volume, given the known value of q_{jk}. Overall then, two dimensions requires one more assumption (approximation) than one dimension. In addition to assuming q is piecewise constant within a control volume (as in one dimension), it is assumed to piecewise constant along the control surface.

A semi-discrete version of the second law can be constructed in exactly the same way that (10.1.8) was constructed. This is

$$(\dot{P}_s)_{jk} \equiv (S_{jk})_{,t} + \frac{a_{j+\frac{1}{2}\,k}}{v_{jk}}\mathbf{F}_{j+\frac{1}{2}\,k}\cdot\mathbf{n} + \frac{a_{j-\frac{1}{2}\,k}}{v_{jk}}\mathbf{F}_{j-\frac{1}{2}\,k}\cdot\mathbf{n} + \frac{a_{j\,k+\frac{1}{2}}}{v_{jk}}\mathbf{F}_{j\,k+\frac{1}{2}}\cdot\mathbf{n} + \frac{a_{j\,k-\frac{1}{2}}}{v_{jk}}\mathbf{F}_{j\,k-\frac{1}{2}}\cdot\mathbf{n} \geq 0 \tag{10.1.12}$$

This does not involve any further assumptions. Since q is assumed piecewise constant along the cell boundary, $\mathbf{F}$ is computed as easily as $\mathbf{f}$. For rectangular meshes this also simplifies. Using the definitions

$$F \equiv \mathbf{F} \cdot \mathbf{i} \qquad \text{and} \qquad G \equiv \mathbf{F} \cdot \mathbf{j} \tag{10.1.13}$$

inequality (10.1.12) can be written, for rectangular meshes, as

$$(\dot{P}_s)_{jk} = (S_{jk})_{,t} + \frac{F_{j+\frac{1}{2}k} - F_{j-\frac{1}{2}k}}{\Delta x} + \frac{G_{jk+\frac{1}{2}} - G_{jk-\frac{1}{2}}}{\Delta y} \geq 0 \tag{10.1.14}$$

which is analogous to (3.1.8). The time derivative can be eliminated as before by using the chain rule. In particular $S_{,t} = S_{,q}{}^T q_{,t}$, where $q_{,t}$ appears in (10.1.10). Making the substitutions yields

$$\begin{aligned}(\dot{P}_s)_{jk} = -(S_{jk})_{,q}{}^T \left(\frac{f_{j+\frac{1}{2}k} - f_{j-\frac{1}{2}k}}{\Delta x} + \frac{g_{jk+\frac{1}{2}} - g_{jk-\frac{1}{2}}}{\Delta y} \right) \\ + \frac{F_{j+\frac{1}{2}k} - F_{j-\frac{1}{2}k}}{\Delta x} + \frac{G_{jk+\frac{1}{2}} - G_{jk-\frac{1}{2}}}{\Delta y} \geq 0\end{aligned} \tag{10.1.15}$$

This is algebraically equivalent to

$$\begin{aligned}(\dot{P}_s)_{jk} = \frac{1}{\Delta x} \Big[&- (S_{jk})_{,q}{}^T (f_{j+\frac{1}{2}k} - f_{jk}) + (F_{j+\frac{1}{2}k} - F_{jk}) \\ &- (S_{jk})_{,q}{}^T (f_{jk} - f_{j-\frac{1}{2}k}) + (F_{jk} - F_{j-\frac{1}{2}k}) \Big] \\ + \frac{1}{\Delta y} \Big[&- (S_{jk})_{,q}{}^T (g_{jk+\frac{1}{2}} - g_{jk}) + (G_{jk+\frac{1}{2}} - G_{jk}) \\ &- (S_{jk})_{,q}{}^T (g_{jk} - g_{jk-\frac{1}{2}}) + (G_{jk} - G_{jk-\frac{1}{2}}) \Big] \geq 0\end{aligned} \tag{10.1.16}$$

which is similar in form to (3.4.3) but with the metric terms left in. This is more easily manipulated in the form

$$(\dot{P}_s)_{jk} = (\dot{P}_s)_{j+\frac{1}{2}k} + (\dot{P}_s)_{j-\frac{1}{2}k} + (\dot{P}_s)_{jk+\frac{1}{2}} + (\dot{P}_s)_{jk-\frac{1}{2}} \geq 0 \tag{10.1.17}$$

where, for example

$$(\dot{P}_s)_{j+\frac{1}{2}k} \equiv \frac{1}{\Delta x} \left[-(S_{jk})_{,q}{}^T (f_{j+\frac{1}{2}k} - f_{jk}) + (F_{j+\frac{1}{2}k} - F_{jk}) \right] \tag{10.1.18}$$

Following the development in one dimension, a cell entropy inequality can be satisfied in several different ways. If the simplicity of a first-order scheme is desired, then

$q_{j+\frac{1}{2}k}, q_{j-\frac{1}{2}k}, q_{jk+\frac{1}{2}}$, and $q_{jk-\frac{1}{2}}$ must each be adjusted in such a way that the individual terms of (10.1.17) are nonnegative. The usual approach of dimensional splitting is equivalent to a pair of constraints.

$$(\dot{P}_s)_{j+\frac{1}{2}k} + (\dot{P}_s)_{j-\frac{1}{2}k} \geq 0 \qquad \text{and} \qquad (\dot{P}_s)_{jk+\frac{1}{2}} + (\dot{P}_s)_{jk-\frac{1}{2}} \geq 0 \tag{10.1.19}$$

The first constraint corresponds to the x direction and the second corresponds to the y direction. These can be satisfied using the one-dimensional techniques already shown. Such an approach will be second-order accurate in each of the coordinate directions, but may have serious anisotropic properties which are generally undesirable.

In fact, it is only the sum $(\dot{P}_s)_{jk}$ which is physically constrained to be nonnegative. This constraint can be met by using a two-dimensional analogue of the one-dimensional technique. First, some second-order accurate states such as $\overline{q}_{j+\frac{1}{2}k}$ are chosen, as was done in one dimension. This can be done by linear averaging, perhaps in the entropy variables. This defines the bar terms such as $\overline{(\dot{P}_s)}_{j+\frac{1}{2}k}$. Summing these gives

$$\overline{(\dot{P}_s)}_{jk} = \overline{(\dot{P}_s)}_{j+\frac{1}{2}k} + \overline{(\dot{P}_s)}_{j-\frac{1}{2}k} + \overline{(\dot{P}_s)}_{jk+\frac{1}{2}} + \overline{(\dot{P}_s)}_{jk-\frac{1}{2}} \geq 0 \tag{10.1.20}$$

If $\overline{(\dot{P}_s)}_{jk} \geq 0$, then (10.1.17) is met by the average states and no adjustment is necessary. If $\overline{(\dot{P}_s)}_{jk} < 0$ there must be one or more negative terms. Each can be reduced in magnitude to ensure $\overline{(\dot{P}_s)}_{jk} \geq 0$. Let the sum of the negative terms be $(\dot{P}_s)^-_{jk}$. This sum must be reduced in magnitude (multiplied) by some positive constant α, $0 \leq \alpha \leq 1$. This constant is given by

$$\alpha = \frac{\overline{(\dot{P}_s)}_{jk}}{(\dot{P}_s)^-_{jk}} \tag{10.1.21}$$

The total change can be divided amongst the negative terms, with those of larger magnitude getting a larger share of the change. For example, if $\overline{(\dot{P}_s)}_{j+\frac{1}{2}k} < 0$ then $q_{j+\frac{1}{2}k}$ needs to be changed in such a way that

$$(\dot{P}_s)_{j+\frac{1}{2}k} \geq \alpha\beta\overline{(\dot{P}_s)}_{j+\frac{1}{2}k}^{\,2} \tag{10.1.22}$$

where

$$\beta \equiv \frac{(\dot{P}_s)^-_{jk}}{(\dot{P}_s^2)^-_{jk}} \tag{10.1.23}$$

and $(\dot{P}_s^2)^-_{jk}$ is the sum of the squares of the negative terms. The desired changes to intermediate states such as $q_{j+\frac{1}{2}k}$ can be accomplished by gradient methods already introduced.

An example suffices to show that this can be a significant improvement over dimensional splitting. Suppose that $\overline{(\dot{P}_s)}_{jk} \geq 0$, but $\overline{(\dot{P}_s)}_{j+\frac{1}{2}k} + \overline{(\dot{P}_s)}_{j-\frac{1}{2}k} < 0$. In this case no

dissipation needs to be added because (10.1.16) is already satisfied. A dimensionally split algorithm would add dissipation anyway because the first inequality of (10.1.19) is not satisfied. In fact, dimensional splitting always requires at least as much dissipation as the more general method of (10.1.22).

10.2 Unstructured Grids

There has been considerable interest recently in unstructured grids. Such grids are substantially easier to generate about complicated geometries than grids which are topologically rectangular. It is difficult to define a suitable dissipation operator on an unstructured grid with existing methods. For example, dimensional splitting is not an option so the usual definitions of total variation or fourth difference operators do not apply.

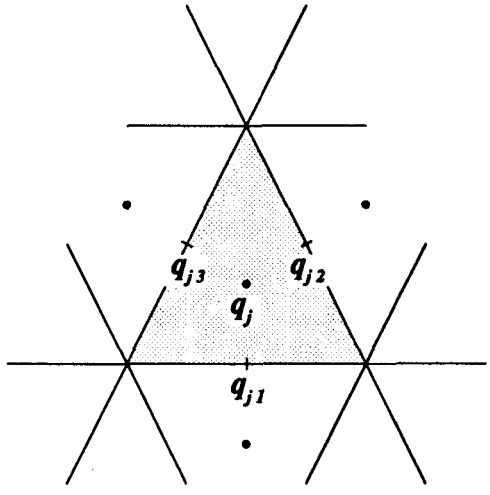

Figure 10.2. – A typical control volume for an unstructured grid.

The simplifications used for rectangular grids do not apply to unstructured grids, one cell of which is shown in figure 10.2. The control volumes are triangles in this case, but they could be quadrilaterals, pentagons or any other polygons. There is no requirement that they be convex or have the same number of sides, as long as they tessellate the domain. The distinguishing feature of unstructured meshes is that there is no obvious numbering scheme, as there is for logically rectangular meshes. This being the case the control volumes are usually identified by a single index, in this case j. Each control volume has a certain number of neighbors (three in this case). The control surface between the i^{th} and j^{th} control volume is denoted by the subscript ij. In figure 10.2, the neighbors of cell j are cells 1,2, and 3 for notational convenience.

The derivation in Section 10.1 can be followed up to equation (10.1.6). At this point it is useful to note that

$$\oint_{\partial V_j} \mathbf{f}_j \cdot \mathbf{n}\, dA = 0 \tag{10.2.1}$$

because the divergence of a constant vector vanishes. Adjusting for the difference in notation and subtracting (10.2.1), (10.1.8) can be written

$$(q_j)_{,t} + \frac{a_{j1}}{v_j}(\mathbf{f}_{j1} - \mathbf{f}_j)\cdot\mathbf{n} + \frac{a_{j2}}{v_j}(\mathbf{f}_{j2} - \mathbf{f}_j)\cdot\mathbf{n} + \frac{a_{j3}}{v_j}(\mathbf{f}_{j3} - \mathbf{f}_j)\cdot\mathbf{n} = 0 \tag{10.2.2}$$

The dot product is carried out for each term, leading to the definitions

$$\begin{aligned} \Delta f_{j1} &\equiv a_{j1}(\mathbf{f}_{j1} - \mathbf{f}_j)\cdot\mathbf{n} \\ \Delta f_{j2} &\equiv a_{j2}(\mathbf{f}_{j2} - \mathbf{f}_j)\cdot\mathbf{n} \\ \Delta f_{j3} &\equiv a_{j3}(\mathbf{f}_{j3} - \mathbf{f}_j)\cdot\mathbf{n} \end{aligned} \tag{10.2.3}$$

With these definitions (10.2.2) simplifies to

$$(q_j)_{,t} + \frac{1}{V_j}[\Delta f_{j1} + \Delta f_{j2} + \Delta f_{j3}] = 0 \tag{10.2.4}$$

In a completely analogous manner, the semi-discrete entropy inequality can be derived.

$$(\dot{P}_s)_j \equiv (S_j)_{,t} + \frac{1}{V_j}[\Delta F_{j1} + \Delta F_{j2} + \Delta F_{j3}] \geq 0 \tag{10.2.5}$$

As before, q_j is assumed to be constant over the control volume and q_{ij} is assumed to be constant over each edge. Using the chain rule again ($S_{,t} = S_{,q}q_{,t}$), it is possible to eliminate $(S_j)_{,t}$ using (10.2.2). This gives, for unstructured meshes, an expression corresponding to (10.1.16).

$$v_j(\dot{P}_s)_j \equiv -(S_{,q})_j[\Delta f_{j1} + \Delta f_{j2} + \Delta f_{j3}] + [\Delta F_{j1} + \Delta F_{j2} + \Delta F_{j3}] \geq 0 \tag{10.2.6}$$

Combining these as before gives

$$v_j(\dot{P}_s)_j \equiv [-(S_{,q})_j\Delta f_{j1} + \Delta F_{j1}] + [-(S_{,q})_j\Delta f_{j2} + \Delta F_{j2}] + [-(S_{,q})_j\Delta f_{j3} + \Delta F_{j3}] \tag{10.2.7}$$

Finally

$$(\dot{P}_s)_j \equiv (\dot{P}_s)_{j1} + (\dot{P}_s)_{j2} + (\dot{P}_s)_{j3} \geq 0 \tag{10.2.8}$$

where, for example

$$(\dot{P}_s)_{j1} \equiv \frac{1}{V_j}[-(S_{,q})_j\Delta f_{j1} + \Delta F_{j1}] \tag{10.2.9}$$

The remainder of the derivation is identical to that of Section 10.1, except that the number of terms in this case is three instead of four. The intermediate states q_{j1}, q_{j2}, and q_{j3} can be constructed so as to satisfy (10.2.8).

It is possible to recover the expressions for regular grids by using the expressions for unstructured grids and including the known areas, volumes, and cell face orientations. Other types of grid systems such as generalized curvilinear grids may also be viewed as specializations of the unstructured grid approach.

10.3 A Multigrid Scheme

Implicit methods as currently practiced require dimensional splitting (approximate factorization) to be practical. For this reason, the unstructured meshes of the previous section, are not conducive to implicit methods. Because of CFL restrictions, explicit methods can be prohibitively expensive. An alternative of current interest is the use of multigrid methods.

If any one feature characterizes multigrid schemes, it is the way they capture the gross features of a solution using a relatively coarse grid, using the usual fine grid to sharpen the solution. This approach is economical for two reasons. First, the coarse grid, having fewer unknowns, requires less arithmetic in each time step. Second, and just as important, those time steps can be bigger for a given CFL number. Looked at another way, if shocks can only move one grid point per step (a common restriction on explicit schemes), it helps to have fewer grid points to traverse.

The second law provides an interesting insight into the stability of certain multigrid schemes. In particular, given a fine grid scheme which satisfies a semi-discrete cell entropy inequality, it is possible to construct multigrid schemes which also do. These multigrid schemes, termed unigrid schemes (ref. 39), were first introduced by Steve McCormick.

In a unigrid scheme, each coarse grid cell consists of a collection of adjacent fine grid cells. Coarser grids contain more fine grid cells, but all grids are composed of a collection of cells from the finest grid. Consider the situation shown in figure 10.3. A single coarse grid cell is represented, which consists of four fine grid cells.

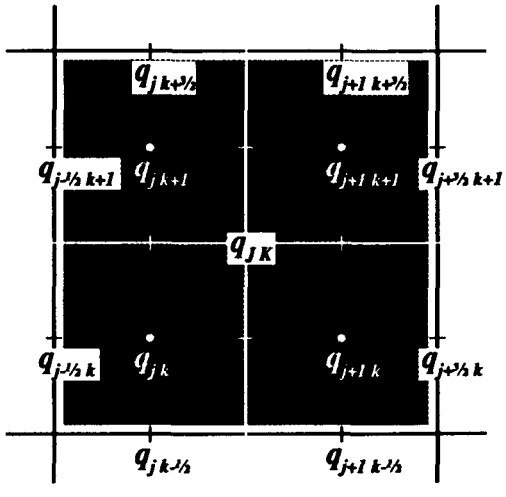

Figure 10.3. – A coarse grid control volume.

To advance in time on the fine grid is not difficult. Section 10.1 showed that schemes which satisfy a cell entropy inequality can be constructed. To advance in time on the coarse grid, consider the coarse grid cell as a single control volume. With this view, it is possible to advance in time without additional assumptions. The surface integral of the fluxes can be computed as on the fine grid using the fact that q is piecewise constant along the edges. In figure 10.3 the surface integral would be done in eight parts.

For the moment, define q_{JK} as the average value over the coarse grid cell. The semi-discrete conservation laws for cell JK can be written

$$\begin{aligned}(q_{JK})_{,t} + \frac{1}{2\Delta x}(f_{j+\frac{3}{2}k+1} + f_{j+\frac{3}{2}k} - f_{j-\frac{1}{2}k+1} - f_{j-\frac{1}{2}k}) \\ + \frac{1}{2\Delta y}(f_{jk+\frac{3}{2}} + f_{j+1k+\frac{3}{2}} - f_{jk-\frac{1}{2}} - f_{j+1k-\frac{1}{2}}) = 0\end{aligned} \tag{10.3.1}$$

where Δx and Δy refer to the fine grid spacing. The problem which remains is to apportion this change among the four cells which make up the larger control volume. The natural choice is to apportion the change according to the volume of each cell. This means

$$(q_{JK})_{,t} = (\overline{q}_{jk})_{,t} = (\overline{q}_{jk+1})_{,t} = (\overline{q}_{j+1k})_{,t} = (\overline{q}_{j+1k+1})_{,t} \tag{10.3.2}$$

This choice may or may not satisfy the second law for the larger control volume. The second law may be written as

$$\begin{aligned}(S_{JK})_{,t}\, v_{JK} + \Delta y(F_{j+\frac{3}{2}k+1} + F_{j+\frac{3}{2}k} - F_{j-\frac{1}{2}k+1} - F_{j-\frac{1}{2}k}) \\ + \Delta x(F_{jk+\frac{3}{2}} + F_{j+1k+\frac{3}{2}} - F_{jk-\frac{1}{2}} - F_{j+1k-\frac{1}{2}}) \geq 0\end{aligned} \tag{10.3.3}$$

The first term, $(S_{JK})_{,t}\, v_{JK}$ represents an integration, carried out as follows:

$$\begin{aligned}(S_{JK})_{,t} v_{JK} =& (S_{jk})_{,t} v_{jk} + (S_{jk+1})_{,t} v_{jk+1} \\ &+ (S_{j+1k})_{,t} v_{j+1k} + (S_{j+1k+1})_{,t} v_{j+1k+1}\end{aligned} \tag{10.3.4}$$

where v_{JK} is the volume of the coarse grid cell and v_{jk}, v_{jk+1}, v_{j+1k}, and v_{j+1k+1} are the volumes of the four fine grid cells. In this case all four are the same, just $\Delta x \Delta y$. The flux terms of (10.3.4) are simply the net export of entropy by convection, the area weighted sum of the entropy flux around the (in this case) eight segments of the control volume boundary.

If (10.3.3) is not satisfied, the fault cannot lie with the flux terms. This is so because the second law can be satisfied in each cell separately, using the same fluxes. The only remaining suspect is $(S_{JK})_{,t}$ which reflects the way that increases in q are distributed.

The term $(S_{JK})_{,t}$ can be increased by mixing a fraction of the total fluid in each cell. The mixed fraction will have an increased entropy. If the mixed fraction reaches unity, the control volume entropy reaches a maximum. Since entropy production depends linearly on this fraction, it is possible to precisely compute the fraction required. A small amount of additional dissipation occurs when the mixed fluid is returned to the cells, and a cell average is taken to determine the final piecewise constant values.

This idea can be quantified using equations such as

$$q_{jk} = \alpha\, q_{JK} + (1-\alpha)\,\overline{q}_{jk} \tag{10.3.5}$$

where $\overline{q}_{jk}$ is obtained when q_{jk} is updated using (10.3.2), and q_{JK} is the average state in the coarse cell. It is also the maximum entropy state for the coarse grid cell. It is easily verified that (10.3.5), if applied to each fine grid cell with the same value of α, is conservative.

The effect of a mixing operation such as this is to increase the entropy within the coarse grid cell. In each fine grid cell

$$S_{jk} \geq \alpha S(q_{JK}) + (1-\alpha)\overline{S_{jk}} \tag{10.3.6}$$

which can also be written

$$S_{jk} - \overline{S_{jk}} \geq \alpha(S(q_{JK}) - \overline{S_{jk}}) \tag{10.3.7}$$

Summing these contributions from each of the control volumes leads to

$$\begin{aligned} v_{JK}(S_{JK} - \overline{S_{JK}}) \geq \alpha \Big[& v_{jk}(S(q_{JK}) - \overline{S_{jk}}) \\ & + v_{jk+1}(S(q_{JK}) - \overline{S_{jk+1}}) \\ & + v_{j+1k}(S(q_{JK}) - \overline{S_{j+1k}}) \\ & + v_{j+1k+1}(S(q_{JK}) - \overline{S_{j+1k+1}}) \Big] \end{aligned} \tag{10.3.8}$$

where S_{JK} is the volume-weighted average of the entropy in the four fine grid cells, and $\overline{S_{JK}}$ is S_{JK} when $(q_{JK})_{,t}$ is distributed according to (10.3.2). The value of S_{JK} that is required can be deduced from the second law. This leaves α as the only unknown. Treating (10.3.8) as an equality, one can solve for α. This value can then be used in (10.3.5) and the resulting distribution will satisfy a (coarse) cell entropy inequality. Following this procedure ensures that, coarse grid or fine, the domain entropy will not decrease. Thus, the stability conjecture of Chapter 2 applies.

As steady state is approached (steady on the scale of the grid at least) the fluxes more nearly balance, and α becomes very small, vanishing at convergence. Away from

convergence however, shocks can propagate much more quickly on the coarse grid than on the fine grid.

There is no particular reason why the coarse grid cells are limited to four fine grid cells. The only requirement is that the coarse grid must use the same volume and flux integrations that the fine grid uses. This ensures that convergence on the finest grid implies convergence on all grids.

10.4 Algorithm Improvements

None of the schemes shown in this work exhibit more than second-order accuracy. There seems to be a need for more accurate schemes, particularly when simulating turbulence without algebraic modeling. This translates into schemes which have less entropy production, for the case of purely convective problems. There are several ways to modify the schemes shown so far, to reduce the cell entropy production rate.

Let's briefly review the second-order schemes shown in Chapters 5, 6, and 7 These methods consist of three parts. First, a "target" value of $\overline{q}_{j+\frac{1}{2}}$ is chosen. Second, the entropy production rate in each cell (or half cell) is computed. Finally the value of $q_{j+\frac{1}{2}}$ is modified as necessary to satisfy a cell entropy inequality. To achieve higher accuracy, one or more of these steps needs to be improved.

Perhaps the simplest thing to change is the target value. Up to now, it has been tacitly assumed that the "best" value for $u_{j+\frac{1}{2}}$ (for scalar equations in one dimension) is the average value, $\frac{1}{2}(u_j + u_{j+1})$. This is the best from a Taylor series viewpoint, but there are other viewpoints.

Recall that

$$(\dot{P}_s)_j = (\dot{P}_s^+)_j + (\dot{P}_s^-)_j \tag{10.4.1}$$

One constraint which is fairly easy to satisfy (in principle) is

$$(\dot{P}_s^+)_j = -(\dot{P}_s^-)_{j+1} \tag{10.4.2}$$

Both sides of the equation depend only on $u_{j+\frac{1}{2}}$. If $u_{j+\frac{1}{2}}$ is chosen to satisfy (10.4.2) then the global entropy production rate will vanish, as for Tadmor's scheme. As discussed in Section 9.3 and 9.4, this property guarantees stability but can lead to unphysical solutions. Deviations from the target value will probably be necessary to satisfy a local entropy inequality. On the other hand, as the mesh becomes finer and finer, these deviations become smaller and smaller (for smooth solutions). This means that unphysical global entropy production vanishes in the limit as the mesh is refined, a desirable property.

For illustration, consider the linear wave equation again. It happens that the value of $u_{j+\frac{1}{2}}$ which satisfies (10.4.2) is the average value. If the target value is used without

modification, the resulting scheme is algebraically equivalent to central differencing with a Crank-Nicolson time advance. Such a scheme is known to be unconditionally stable but far from TVD, generating severe oscillations.

When the values of $u_{j+\frac{1}{2}}$ are modified to satisfy a cell entropy inequality as well, the resulting scheme is the second-order accurate one discussed in Section 5.7. Typical solutions are shown in figure 5.8. Thus, (10.4.2) has already been tried for the wave equation. It has not led to any schemes with better than second-order accuracy. While simple and elegant, (10.4.2) is not enough by itself.

Another way to pick target values is to include more points in the stencil. Perhaps the ultimate in schemes of this type is the subcell resolution method of arbitrary order proposed by Harten (ref. 40), and Harten, Enquist, and Osher mref41. Although such schemes are essentially nonoscillatory (ENO), they are known to contain small violations of local entropy conditions. The gradient following techniques discussed in Section 5 show how to modify $u_{j+\frac{1}{2}}$ so as to correct these small defects. Unfortunately, the ENO approach is restricted to one dimension.

Regardless of which target value is selected, the method outlined in Section 5.3 can be used to modify it in such a way that a cell entropy inequality is satisfied. Since this involves finding a root of a nonlinear equation, iteration is generally called for. Unfortunately, after all the work of solving for ϕ to several decimal places, the resulting dissipation is usually about twice what is required (see fig. 7.2). This is because the increase in $(\dot{P}_s^+)_j$ is about equal to the increase in $(\dot{P}_s^-)_{j+1}$ (at $\phi = 1$ the slopes are equal if entropy variables are used). To first-order, the increase in $(\dot{P}_s^-)_j$ will be equal to the increase in $(\dot{P}_s^+)_j$. In the worst case, however, $(\dot{P}_s^-)_j$ doesn't change, and $(\dot{P}_s^+)_j$ must allow for this. The double correction occurs because of this disparity between the worst case and the usual case.

This suggests under-relaxing the first nonlinear iteration by a factor of two and recalculating $(\dot{P}_s)_j$. After this iteration $(\dot{P}_s)_j$ will be very close to zero in most the cells in which it had been negative. It is recommended that this underrelaxation be discontinued after one iteration.

A third way to reduce entropy production is to extend the method suggested in Section 5.3 in the following way. After $u_{j+\frac{1}{2}}$ has been modified to make $(\dot{P}_s)_j > 0$ in each cell, $(\dot{P}_s)_j$ can be reevaluated. This value can be used to give each half cell an "entropy budget" with which to move $\phi_{j+\frac{1}{2}}$ back toward the target value, thereby reducing the entropy production in cells j and $j+1$. The variations on this third approach are nearly infinite, but they are bound to be costly.

10.5 Grid Refinement

One of the things which makes CFD an art instead of a science is the construction of a suitable computational mesh. It is often very difficult to construct a regular mesh about a complicated geometry. Small changes in geometry entail large changes in such grids. As the complexity increases, control over placement of grid points decreases dramatically. By contrast, construction of an unstructured mesh is not at all difficult, especially if mesh clustering is not an issue.

A promising approach is to produce a solution dependent clustering by moving the existing grid points (ref. 42), or by adding new ones (ref. 43). With these two techniques, a computational mesh can be modified to improve resolution of steady or transient flow features of interest. Some researchers have gone so far as to generate a new mesh every time step (ref. 44).

Two problems have always hampered this approach. The first is the selection of a suitable sensor for detecting underresolved and over resolved areas. The second problem is the tendency to place all the points in the shock, which in the inviscid case is a true discontinuity and therefore unresolvable.

The second law provides an elegant sensor, at least for inviscid problems. Such problems produce no entropy except at shocks. This can be shown analytically. Numerically though, some (hopefully positive) rate of entropy production exists in each cell. Ignoring shocks for the moment, this cell entropy production rate $(\dot{P}_s)_j$ is a measure of resolution. The higher $(\dot{P}_s)_j$ is, the more poorly resolved the solution is locally.

The problem of course is shocks. Shocks have enormous entropy production rates (both physically and numerically) and tend to invalidate the sensor. Standard techniques include imposing a minimum cell size (for mesh movement techniques) or simply refining all cells (including those with shocks) which have more than some threshold entropy production rate (for mesh refinement techniques).

10.6 Navier-Stokes Equations

In many flows of interest, it is important to account for the effects of viscosity and heat transfer. The inclusion of these effects is what separates the Navier-Stokes equations from the Euler equations. This section will explore the effects of viscosity and heat transfer on the second law.

The viscous terms are stresses, forces that act on the boundaries of each control volume. These can affect the momentum and kinetic energy in the volume, which in turn influence the rate of entropy accumulation, $S_{,t}$. They do not, however, directly affect the entropy flux F because they do not change the state of the fluid at the cell boundary.

The major effect of viscous stresses acting on a cell boundary is to reduce velocity differences between adjacent cells. Since momentum is conserved, this implies a reduction in the total kinetic energy of the two adjacent cells. Since energy is conserved the lost kinetic energy must appear as heat. The process of converting kinetic energy into heat produces entropy.

The heat transfer terms appear in the energy equation and thereby influence the rate of entropy accumulation $S_{,t}$. They are part of the definition of entropy flux. In particular, the term $-\frac{\dot{Q}}{T}$ is added to the entropy flux F. The heat flux $\dot{Q}$ is usually determined by Fourier's law; it is proportional to (but of opposite sign to) the temperature gradient.

It is well known that entropy is always produced by heat transfer over a finite temperature difference. Given a heat flux $\dot{Q}$ between two cells of different temperatures, the entropy flux leaving the high temperature cell is $\frac{\dot{Q}}{T_h}$ while the entropy flux entering the low temperature cell is $\frac{\dot{Q}}{T_l}$. Both temperatures are measured in degrees Kelvin or some other absolute temperature scale. Since the heat flux is generally from the hot cell to the cold cell, the entropy of the cold cell increases by more than the entropy of the hot cell decreases. This difference is the entropy production due to heat transfer.

The inclusion of viscosity and thermal conductivity has interesting consequences. For example, since some entropy is produced "legitimately," it becomes possible to scale back the "artificial" dissipation because this was added only for stability reasons. In some cases with very fine grids or very low Reynolds numbers, it may be possible to do away with it altogether. An extreme example is the heat equation for which no artificial dissipation is required.

Secondly, including these terms may prove to be a far easier way to introduce entropy production than the way shown in Chapter 7. The Reynolds number and Prandtl number can be used as locally adjustable parameters which are set in some way which satisfies the cell entropy inequality. Similar ideas have been proposed by Dulikravitch (ref. 45).

10.7 Chemistry

Recently there has been some interest in computing very high speed flows ($M > 10$)[1] At the temperatures involved, oxygen and nitrogen react in ways that would be extremely unlikely at room temperature. This necessitates a different sort of entropy accounting.

The first major change is that the convective entropy flux (as opposed to the heat transfer term) must be computed separately for each species being tracked. That is, the

[1] It is debatable whether such flows can be described by a continuum approximation at all since the Knudsen number is too big. The mean free path can exceed the boundary layer thickness in some cases. This issue is ignored here.

entropy carried by a molecule of oxygen is different than that of a molecule of nitrogen. It is no longer enough to know the mass transferred, the mass flux must be known for each species individually.

The second major effect is that reactions occur within the control volume at a rate that depends on the concentrations of the reactants in the control volume. These reactions produce entropy. This must be reflected in the integration of the entropy within each cell volume. Each species must be accounted for separately as in the flux computation.

10.8 Stability and Unsteady Flow

The stability analysis presented here is more suitable for unsteady flows than is the more traditional von Neumann analysis. It is also more suitable for the nonlinear dissipation techniques currently practiced.

Traditional von Neumann analysis is fairly simple and therefore quite popular. It consists of supposing that the solution consists of a sine wave, and then applying the numerical scheme to this solution. If it can be shown that the solution will not increase in amplitude, regardless of frequency, then we can say with confidence that the scheme is stable in sense of Lax. The L_2 norm in wave space is bounded for all time.

There are several problems with this. First, it is generally difficult to apply von Neumann analysis to nonlinear numerical schemes. By their nature, nonlinear schemes do not maintain the frequency of a sine wave. Even if a single frequency is used as an initial condition, the solution after one step contains many frequencies. The use of nonlinear dissipation is important. For example, it is impossible to construct a TVD scheme (even for one-dimensional wave equations, where this property is appropriate) which is more than first-order accurate, without using nonlinear dissipation (ref. 29).

In many physical flows, disturbances of certain frequencies actually grow. This is resolved ultimately by transfer of energy into frequencies which don't grow. Von Neumann stability analysis, by requiring stability in each mode separately, does not allow accurate simulation of such flows.

By contrast, the satisfaction of a cell entropy inequality is no harder for nonlinear schemes than for linear schemes. The transfer of energy between frequencies does not affect the analysis. In addition, there are no physically allowable flows which violate the second law. Thus, there is no danger of excluding an interesting physical flow on the grounds that it is unstable.

10.9 Summary

There is considerable distance between the proposals of this chapter and the reality of a working simulation. (This is why this dissertation contains no results in higher di-

mensions or for more complicated equation sets.) Nevertheless, the possibilities appear bright; much more can be done than was previously possible. Provided that the conjecture of Chapter 2 holds, Chapter 7 demonstrates a scheme with true nonlinear stability for the Euler equations in one dimension. It appears that such schemes are extendable to multidimensions without dimensional splitting. Extension to the Navier-Stokes equations also seems possible.

The basic principle, a cell entropy inequality, appears sound for any problem in continuum mechanics. By contrast, standard upwind schemes are limited to equations which have real eigenvalues and linearly independent eigenvectors. For practical purposes the eigensystem must be expressed in closed form. This may prove to be a serious limitation.

Many schemes in use today rely on a geometric or intuitive definition of smoothness. These are subject to numerous pitfalls, the most serious of which is the lack of any connection to the physics. Many of them are also limited to one-dimensional representations. The cell entropy inequality approach does not share these pitfalls.

Where the upwind methods shine is in one dimension. Here the *de facto* similarity between schemes which satisfy a cell entropy inequality and upwind schemes already in use is striking. From this point of view, the second law is merely a way of generalizing the schemes already in use. This, and the fact that there are no obvious roadblocks to the really interesting problems, leads to the conclusion that the ideas presented here will provide fertile ground for future research.

Chapter 11. SUMMARY

In the author's view, CFD is presently an art with elements of science. There are several steps in which "a miracle occurs." A particular strategy is advocated without good reason, simply because it seems to work and nothing more satisfying is available. This work has addressed one such area, that of artificial dissipation. Artificial dissipation is analyzed from the viewpoint of numerical satisfaction of the second law of thermodynamics.

A strong conjecture of nonlinear stability is offered for schemes which satisfy the second law over the entire domain (the cell-by-cell entropy inequality is not required for this). This conjecture is currently limited to periodic boundary conditions.

Subsequent investigation provided indications that a cell-by-cell application of the second law provides more satisfactory solutions. Part of the difficulty in providing a proof lies in defining what is meant by "satisfactory." For example, consider solution of the scalar wave equation, using three-point central differencing and Crank-Nicolson time advance. Such a scheme produces no net entropy over the domain, the physically correct result. It does, however, have a negative entropy production rate in some cells and a positive rate in others. The solution is bounded as required by the proof (and by linear theory), but it contains numerous oscillations which are not physically correct. On the other hand, if this scheme is modified slightly to require a positive entropy production rate in each cell, the oscillations do not appear. Similar observations were made for a nonlinear, entropy conserving scheme in Chapter 9. There it was noted that the wiggles were correlated in time, space, and magnitude with the negative entropy production rate in certain cells.

The schemes suggested by this approach were formally shown to have the property of diminishing total variation as defined by Harten when applied to the linear wave equation. In the case of the nonlinear wave equation, the TVD property could be achieved with the assumption of locally constant wave speed. For the Euler equations in one dimension, there are still some similarities, but one major difference appears: while it isn't possible to construct a scheme which is TVD for this problem, it is possible to satisfy a cell entropy inequality.

Satisfaction of a cell entropy inequality for the one-dimensional Euler equations was facilitated by the use of a surprising identity (7.3.1). This was discovered using MACSYMA, not derived with great insight. The author is not aware of any previous mention of this identity in the literature. It was pointed out by Roe that this identity also applies to the shallow water equations. Such an identity provides an unambiguous scaling of the eigenvectors in such problems. This is helpful in theoretical work but not too important in practical schemes.

Various extensions appear possible including multidimensions, viscosity, heat transfer, and chemistry. This flexibility is possible because of the widespread applicability of the second law. Since dimensional splitting is not required, unstructured grids become feasible. These may be refined using the entropy production rate itself as an indicator of local resolution.

In addition to artificial dissipation and grid refinement, topics covered in this work, there are several more "black boxes" which are used in CFD. These are turbulence modeling (or turbulence itself!), boundary conditions, and nonlinear accuracy estimates. Until these are also understood, Computational Fluid Dynamics will remain on a house of cards, depending for validation on the very experiments it is supposed to render unnecessary.

Appendix A. DERIVATION OF SUFFICIENT CONDITIONS FOR TVD SCHEMES

The objective here is to show that the total variation of the solution will not increase if the sufficient conditions are met.

The derivation begins with the assumption that the independent variable is a scalar quantity u and that it can be advanced in time on a discrete mesh using

$$u_j^{n+1} = u_j^n - C_{j-\frac{1}{2}}\Delta u_{j-\frac{1}{2}} + D_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}} \tag{A.1.1}$$

where $C_{j-\frac{1}{2}}$ and $D_{j+\frac{1}{2}}$ can be functions of u at the grid points.

The objective is to derive conditions such that

$$\mathrm{TV}(u^{n+1}) \leq \mathrm{TV}(u^n) \tag{A.1.2}$$

where TV is the so-called total variation of u. Total variation is defined as a sum so that (A.1.2) can be written as

$$\sum_j |u_{j+1}^{n+1} - u_j^{n+1}| \leq \sum_j |u_{j+1}^n - u_j^n| \tag{A.1.3}$$

The sum is carried out over all values of j, that is, over all mesh intervals.

The next step is to eliminate those variables which are at time level $n+1$. Using (A.1.1) it is possible to determine u_{j+1}^{n+1} by a simple index shift.

$$u_{j+1}^{n+1} = u_{j+1}^n - C_{j+\frac{1}{2}}\Delta u_{j+\frac{1}{2}} + D_{j+\frac{3}{2}}\Delta u_{j+\frac{3}{2}} \tag{A.1.4}$$

Subtracting (A.1.1) gives an expression for $u_{j+1}^{n+1} - u_j^{n+1}$.

$$\begin{aligned} u_{j+1}^{n+1} - u_j^{n+1} =& (u_{j+1}^n - u_j^n) + C_{j-\frac{1}{2}}\Delta u_{j-\frac{1}{2}} \\ & - (C_{j+\frac{1}{2}} + D_{j+\frac{1}{2}})\Delta u_{j+\frac{1}{2}} + D_{j+\frac{3}{2}}\Delta u_{j+\frac{3}{2}} \end{aligned} \tag{A.1.5}$$

Collecting terms gives

$$u_{j+1}^{n+1} - u_j^{n+1} = C_{j-\frac{1}{2}}\Delta u_{j-\frac{1}{2}} + (1 - C_{j+\frac{1}{2}} - D_{j+\frac{1}{2}})\Delta u_{j+\frac{1}{2}} + D_{j+\frac{3}{2}}\Delta u_{j+\frac{3}{2}} \tag{A.1.6}$$

At this point one takes the absolute value of both sides and applies the sum inequality for absolute values

$$|u_{j+1}^{n+1} - u_j^{n+1}| \leq |C_{j-\frac{1}{2}}\Delta u_{j-\frac{1}{2}}| + |(1 - C_{j+\frac{1}{2}} - D_{j+\frac{1}{2}})\Delta u_{j+\frac{1}{2}}| + |D_{j+\frac{3}{2}}\Delta u_{j+\frac{3}{2}}| \tag{A.1.7}$$

Summing both sides,

$$\sum_j |u_{j+1}^{n+1} - u_j^{n+1}| \leq \sum_j \Big(|C_{j-\frac{1}{2}} \Delta u_{j-\frac{1}{2}}| + |(1 - C_{j+\frac{1}{2}} - D_{j+\frac{1}{2}}) \Delta u_{j+\frac{1}{2}}| + |D_{j+\frac{3}{2}} \Delta u_{j+\frac{3}{2}}| \Big) \tag{A.1.8}$$

Splitting the products and rearranging terms gives

$$\begin{aligned} \mathrm{TV}(u^{n+1}) \leq & \sum_j (|C_{j-\frac{1}{2}}||\Delta u_{j-\frac{1}{2}}|) \\ & + \sum_j (|(1 - C_{j+\frac{1}{2}} - D_{j+\frac{1}{2}})||\Delta u_{j+\frac{1}{2}}|) \\ & + \sum_j (|D_{j+\frac{3}{2}}||\Delta u_{j+\frac{3}{2}}|) \end{aligned} \tag{A.1.9}$$

At this point the key observation is made. Notice that, for example

$$\sum_j (|C_{j-\frac{1}{2}}||\Delta u_{j-\frac{1}{2}}|) = \sum_j (|C_{j+\frac{1}{2}}||\Delta u_{j+\frac{1}{2}}|) \tag{A.1.10}$$

except for boundary conditions. The boundary conditions are ignored in this derivation so the indices $j - \frac{1}{2}$ and $j + \frac{3}{2}$ can both be shifted to $j + \frac{1}{2}$. This means that (A.1.9) may be written as

$$\begin{aligned} \mathrm{TV}(u^{n+1}) \leq & \sum_j (|C_{j+\frac{1}{2}}||\Delta u_{j+\frac{1}{2}}|) \\ & + \sum_j (|(1 - C_{j+\frac{1}{2}} - D_{j+\frac{1}{2}})||\Delta u_{j+\frac{1}{2}}|) \\ & + \sum_j (|D_{j+\frac{1}{2}}||\Delta u_{j+\frac{1}{2}}|) \end{aligned} \tag{A.1.10}$$

which allows a massive collection of terms

$$\mathrm{TV}(u^{n+1}) \leq \sum_j (|C_{j+\frac{1}{2}}| + |(1 - C_{j+\frac{1}{2}} - D_{j+\frac{1}{2}})| + |D_{j+\frac{1}{2}}|)|\Delta u_{j+\frac{1}{2}}| \tag{A.1.10}$$

The only thing which prevents further collapse of the right-hand side is the absolute value operator. If the three terms are individually positive the absolute value operator is redundant. That is if

$$\begin{aligned} C_{j+\frac{1}{2}} &\geq 0 \\ D_{j+\frac{1}{2}} &\geq 0 \\ C_{j+\frac{1}{2}} + D_{j+\frac{1}{2}} &\leq 1 \end{aligned} \tag{A.1.11}$$

for all j, then the term in parentheses evaluates to unity and

$$\mathrm{TV}(u^{n+1}) \leq \mathrm{TV}(u^n) \tag{A.1.12}$$

Conditions (A.1.11), first derived by Harten, are known as the sufficient conditions for TVD schemes.

REFERENCES

1. Lax, P.D., *Hyberbolic Systems Of Conservation Laws And The Mathematical Theory Of Shock Waves*, in "CBMS Regional Conference Series In Applied Mathematics," SIAM, Philadelphia, 1973.

2. Murman, E.M. and Cole, J.D., *Calculation Of Plane Steady Transonic Flows*, AIAA Journal V9(1) (1971), 114–121.

3. MacCormack, R.W., *Current Status Of Numerical Solutions Of The Navier Stokes Equations.*, AIAA paper 85–0032 (Jan. 1985), p. 1.

4. Steger, J.L., *Implicit Finite Difference Simulation Of Flow About Arbitrary Geometries With Application To Airfoils*, AIAA paper 77–665 (Jun. 1977).

5. Pulliam, T.H. and Steger, J.L., *Implicit Finite-Difference Simulation Of Three Dimensional Compressible Flow*, AIAA Journal V18(2) (1980), 159–167.

6. Baldwin, B.S. and Lomax, H., *Thin Layer Approximation And Algebraic Model For Separated Turbulent Flows*, AIAA paper 78–257 (Jan. 1978).

7. Osher, S. and Chakravarthy, S., *Upwind Schemes And Boundary Conditions With Applications To Euler Equations In General Geometries*, J. Comp. Phy. V50(3) (1983), 447–481.

8. Roe, P.L., *The Use Of The Riemann Problem In Finite Difference Schemes*, Proceedings of the Seventh International Conference On Numerical Methods In Fluid Dynamics, Stanford, CA (1980), 354–359, Springer-Verlag.

9. Steger, J.L. and Warming, R.F., *Flux Vector Splitting Of The Inviscid Gas Dynamic Equations With Applications To Finite Difference Methods*, J. Comp. Phy. V40(2) (1981), 263–293..

10. Boris, J.P. and Book, D.L., *Flux Corrected Transport, I, SHASTA, A Fluid Transport Algorithm That Works*, J. Comp. Phy. V11(1) (1973), 38–69.

11. Harten, A., *High Resolution Schemes For Hyperbolic Conservation Laws*, J. Comp. Phy. V49(3) (1983), 357–393.

12. Glimm, J. and Lax, P.D., *Decay Of Solutions Of Systems Of Nonlinear Hyperbolic Conservation Laws*, Mem. Amer. Math. Soc. V101 (1970).

13. Oleinik, O.A., *Discontinuous Solutions Of Non-Linear Differential Equations*, Uspekhi Mat. Nauk(N.S.) V12(3) (1957), 3–73; American Math. Soc. Transl., Ser. 2 V26, 95–172.

14. Osher, S., *Riemann Solvers, The Entropy Condition, And Difference Approximations*, SIAM J. Num. Anal. V21 (1984), 217–235.

15. Tadmor, E., *The Numerical Viscosity Of Entropy Stable Schemes For Systems Of Conservation Laws. I*, Mathematics of Computation V49 (1987), 91–103.

16. Hughes, T.J.R.,Franca, L.P., and Mallet, M., *A New Finite Element Formulation For Computational Fluid Dynamics: I. Symmetric Forms Of The Compressible Euler And Navier-Stokes Equations And The Second Law Of Thermodynamics*, Comp. Meth. App. Mech. Eng. V54(2) (1986), 223–235.

17. Reynolds, W.C., "Thermodynamics," Second ed., McGraw Hill Inc., 1965,1968, p. 67.

18. Reynolds, W.C., "Thermodynamics," Second ed., McGraw Hill Inc., 1965,1968, p. 224.

19. Lax, P.D., *Shock Waves And Entropy*, Proceedings of a symposium at the University of Wisconsion, E.H. Zarantonello, ed. (1971), 603–634.

20. Krushkov, S.N., *First Order Quasi-Linear Equations In Several Independent Variables*, Math. USSR Sb. V10(2) (1970), 217–243..

21. von Neumann, J., *Theory Of Shock Waves*, "John von Neumann, Collected Works," Second ed., Pergammon Press, Oxford, England, 1943, pp. 178–202.

22. Harten, A., *On The Symmetric Form Of Systems Of Conservation Laws With Entropy*, J. Comp. Phy. V49(1) (1983), 151–164..

23. Mallet, M., *A Finite Element Method For Computational Fluid Dynamics*, Ph.D. Thesis, Stanford University (1985).

24. Dutt, P., *Stable Boundary Conditions And Difference Schemes For Navier–Stokes Equations*, SIAM J. Num. Anal. V25(2) (1988), 245–267; ICASE report number 85–37 (1985).

25. Merriam, M.L., *Smoothing And The Second Law*, Comp. Meth. App. Mech. Eng. V64(1) (1987), 177–193.

26. Mock, M.S., *Systems Of Conservation Laws Of Mixed Type*, J. Diff. Eq. V37 (1980), 70–88.

27. Sweby, P.K., *High Resolution Schemes Using Flux Limiters For Hyperbolic Conservation Laws*, SIAM J. Num. Anal. V21(5) (1984), 995–1011.

28. Crandall, M.G. and Madja, A., *Monotone Difference Approximations For Scalar Conservation Laws*, Mathematics of Computation V34(149) (1980), 1–21.

29. Godunov, S.K., *A Difference Method For Numerical Calculation Of Discontinuous Solutions Of The Equations Of Hydrodynamics*, Mat Sbornik (N.S.) V47(89) (1959), 271–306.

30. Jameson, A. and Schmidt, W., *Some Recent Developments In Numerical Methods For Transonic Flows*, Comp. Meth. App. Mech. Eng. V51 (1985), 467–493.

31. Pulliam, T.H., *Artificial Dissipation Models For The Euler Equations*, AIAA Journal V24(12) (1986), 1931–1940.

32. Barth, T.J., *Finite Domain Construction Of TVD Schemes*, in "Proceedings of Second International Conference on Hyperbolic Problems, March 14–18," 1988, pp. 9–13.

33. Yee, H.C., *Upwind And Symmetric Shock Capturing Schemes*, NASA TM–89464 (1987).

34. Harten, A., Lax, P.D., and van Leer, B., *On Upstream Differencing And Godunov-Type Schemes For Hyperbolic Conservation Laws*, ICASE report number 82–5 (1982).

35. Yee, H.C., Warming, R.F. and Harten, A., *Implicit Total Variation Diminishing (TVD) Schemes For Steady State Calculations*, J. Comp. Phy. V57 (1985), p. 327.

36. Davis, S.F., *TVD Finite Difference Schemes And Artificial Viscosity*, ICASE report number 84–20 (1984).

37. Yee, H.C., *Construction Of Explicit And Implicit Symmetric TVD Schemes And Their Applications*, J. Comp. Phy. V68(1) (1987), 151–179.

38. Goodman, J.B. and LeVeque, R.J., *On The Accuracy Of Stable Schemes For 2D Conservation Laws*, Mathematics Of Computation V45(171) (1985), 15–21.

39. McCormick, S.F. and Ruge, J.W., *Unigrid For Multigrid Simulation*, Mathematics Of Computation V41 (1983), 43–62.

40. Harten, A., *Eno Schemes With Subcell Resolution*, ICASE Report No. 87–56 (1987.).

41. Harten, A., Engquist, and B., Osher, S., *Uniformly High Order Accurate Essentially Non-Oscillatory Schemes III*, J. Comp. Phy. V71 (1987), 231–303.

42. Nakahashi, K. and Deiwert, G.S., *A Practical Adaptive Grid Method For Complex Fluid Flow Problems.*

43. Dannenhoffer, J.F. III and Baron, J.R., *Grid Adaptation For The 2-D Euler Equations*, AIAA paper 85–0484 (Jan. 1985).

44. Löhner, R., *The Efficient Simulation Of Strongly Unsteady Flows By The Finite Element Method*, AIAA paper 87–0555 (Jan. 1987).

45. Dulikravitch, G.S. and Dorney, D.J., *Numerical Versus Physical Dissipation In The Solution Of Compressible Navier-Stokes Equations*, AIAA paper 89–0550 (Jan. 1989).

NASA
National Aeronautics and Space Administration

Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA TM-101086		

4. Title and Subtitle

An Entropy-Based Approach to Nonlinear Stability

5. Report Date

March 1989

6. Performing Organization Code

7. Author(s)

Marshal L. Merriam

8. Performing Organization Report No.

A-89078

10. Work Unit No.

505-60-01

9. Performing Organization Name and Address

Ames Research Center
Moffett Field, CA 94035

11. Contract or Grant No.

13. Type of Report and Period Covered

Technical Memorandum

12. Sponsoring Agency Name and Address

National Aeronautics and Space Administration
Washington, DC 20546-0001

14. Sponsoring Agency Code

15. Supplementary Notes

Point of Contact: Marshal L. Merriam, Ames Research Center, MS 202A-1, Moffett Field, CA 94035
(415) 694-4737 or FTS 464-4737

16. Abstract

Many numerical methods used in Computational Fluid Dynamics (CFD) incorporate an artificial dissipation term to suppress spurious oscillations and control nonlinear instabilities. The same effect can be accomplished by using upwind techniques, sometimes augmented with limiters to form Total Variation Diminishing (TVD) schemes. An analysis based on numerical satisfaction of the second law of thermodynamics allows many such methods to be compared and improved upon.

A nonlinear stability proof is given for discrete scalar equations arising from a conservation law. Solutions to such equations are bounded in the L_2 norm if the second law of thermodynamics is satisfied in a global sense over a periodic domain. It is conjectured that an analogous statement is true for discrete equations arising from systems of conservation laws.

Analysis and numerical experiments suggest that a more restrictive condition, a positive entropy production rate in each cell, is sufficient to exclude unphysical phenomena such as oscillations and expansion shocks. Construction of schemes which satisfy this condition is demonstrated for linear and nonlinear wave equations and for the one-dimensional Euler equations.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Entropy Stability Total variation diminishing	Unclassified-Unlimited Subject Category - 64

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of pages	22. Price
Unclassified	Unclassified	153	A08

For sale by the National Technical Information Service, Springfield, Virginia 22161