17

Source Terms and Balance Laws

So far, we have only considered homogeneous conservation laws of the form $q_t + f(q)_x = 0$. As mentioned briefly in Section 2.5, there are many situations in which *source terms* also appear in the equations, so that we wish to solve the system

$$q_t + f(q)_x = \psi(q). \tag{17.1}$$

Note that these are generally called source terms even if physically they represent a sink rather than a source (i.e., net loss rather than gain of the quantity q). The equation (17.1) is also often called a *balance law* rather than a conservation law. We start with a few examples showing how source terms can arise. Others will be encountered later.

Reacting Flow

Fluid dynamics problems often involve chemically reacting fluids or gases. An example was mentioned in Section 2.5. An even simpler example is studied below in Section 17.2. In these examples the reacting species are assumed to represent a small fraction of the volume, and the chemical reactions have no effect on the fluid dynamics. More interesting problems arise when the reactions affect the fluid motion, as in combustion problems where the heat released by the reactions has a pronounced effect on the dynamics of the flow. Often the chemical reactions occur on much faster time scales than the fastest wave speeds in the gas, resulting in problems with *stiff source terms*. Some of the issues that arise in this case are discussed in Section 17.10.

External Forces Such as Gravity

In introducing gas dynamics in Section 2.6, we ignored external forces acting on the gas, such as gravity. External forces give source terms in the momentum equation, and we then do not expect conservation of the initial momentum, since this force will lead to an acceleration of the fluid and a change in its net momentum.

As an example, consider the equations of one-dimensional isentropic gas dynamics in the presence of a gravitational field, pointing in the negative x-direction (so x now measures distance above the earth, say, in a column of gas). The gravitational force acting on the gas causes an acceleration, and hence this force enters into the integral equation for the time-derivative of momentum. The equation (2.33) is replaced by

$$\frac{d}{dt}\int_{x_1}^{x_2} \rho(x,t)u(x,t)\,dx = [\rho(x_1,t)u^2(x_1,t)+p(x_1,t)]-[\rho(x_2,t)u^2(x_2,t)+p(x_2,t)] \\ -\int_{x_1}^{x_2} g\rho(x,t)\,dx. \tag{17.2}$$

The differential equation (2.34) becomes

$$(\rho u)_t + (\rho u^2 + p)_x = -g\rho. \tag{17.3}$$

In this case the system (2.39) with q and $f(q)$ given by (2.40) is augmented by a source term with

$$\psi(q) = \begin{bmatrix} 0 \\ -gq^1 \end{bmatrix}.$$

Geometric Source Terms

Often a physical problem in three space dimensions can be reduced to a mathematical problem in one or two dimensions by taking advantage of known symmetries in the solution. For example, if we wish to compute a spherically expanding acoustic wave arising from a pressure perturbation at one point in space, then we can solve a one-dimensional problem in r (distance from the source) and time. However, the homogeneous conservation law in three dimensions may acquire source terms when we reduce the dimension. This follows from the fact that the interval $[r_1, r_2]$ now corresponds to a spherical shell, whose volume varies with r like r^2. Hence a substance whose total mass is fixed but that is spreading out radially will have a density (in mass per unit volume) that is decreasing as the substance spreads out over larger and larger spheres.

As an example, in Section 18.9 we will see that with radial symmetry, the three-dimensional acoustics equations can be reduced to the one-dimensional system

$$\begin{aligned} p_t + K_0 u_r &= -\frac{2K_0 u}{r}, \\ \rho_0 u_t + p_r &= 0, \end{aligned} \tag{17.4}$$

where u is now the velocity in the radial direction r. This has the same form as the one-dimensional equations of acoustics but with a geometric source term. Similar geometric source terms arise if we consider flow in a channel or nozzle with varying cross-sectional area. If the variation is relatively slow, then we may be able to model this with a one-dimensional system of equations that includes source terms to model the area variation. This was discussed for a simple advection equation in Section 9.1. More generally this leads to *quasi-one-dimensional* models.

Higher-Order Derivatives

Our focus is on developing methods for first-order hyperbolic systems, but many practical problems also involve higher-order derivatives such as small viscous or diffusive terms. Examples include the advection–diffusion equation $q_t + \bar{u}q_x = \mu q_{xx}$, the Navier–Stokes equations in which viscous terms are added to the Euler equations, or the shallow water equations with bottom friction. Other problems may instead (or in addition) include dispersive terms involving q_{xxx} or other odd-order derivatives. An example is the *Korteweg–de Vries* (KdV) *equation* $q_t + qq_x = q_{xxx}$. Then the equations may be of the form (17.1) with $\psi(q)$ replaced by $\psi(q, q_{xx}, q_{xxx}, \dots)$. We can still view ψ as a source term and apply some of the techniques developed in this chapter. In particular, fractional-step methods are often used to incorporate viscous terms in fluid dynamics problems. This is discussed briefly in the next section and further in Section 17.7.

17.1 Fractional-Step Methods

We will primarily study problems where the homogeneous equation

$$q_t + f(q)_x = 0 \tag{17.5}$$

is hyperbolic and the source terms depend only on q (and perhaps on x) but not on derivatives of q. In this case the equations

$$q_t = \psi(q) \tag{17.6}$$

reduce to independent systems of ODEs at each point x.

One standard approach for such problems is to use a *fractional-step* or *operator-splitting* method, in which we somehow alternate between solving the simpler problems (17.5) and (17.6) in order to approximate the solution to the full problem (17.1). This approach is quite simple to use and is implemented in the CLAWPACK software (see Section 5.4.6). It allows us to use high-resolution methods for (17.5) without change, coupling these methods with standard ODE solvers for the equations (17.6). This approach is described in more detail and analyzed in this chapter. There are situations where a fractional-step method is not adequate, and the analysis presented in this chapter will shed some light on the errors this splitting introduces and when it can be successfully used.

There are also many situations in which the hyperbolic equation is coupled with other terms that involve derivatives of q. For example, the advection–diffusion equation

$$q_t + \bar{u}q_x = \mu q_{xx}$$

can be viewed as an equation of the general form (17.1) in which ψ depends on q_{xx}. This equation should more properly be viewed as

$$q_t + (\bar{u}q - \mu q_x)_x = 0,$$

in conservation form with the flux function $\bar{u}q - \mu q_x$ (see Section 2.2). But in practice it is often simplest and most efficient to use high-resolution explicit methods for the advection part and implicit methods for the diffusion equation $q_t = \mu q_{xx}$, such as the Crank–Nicolson method (4.13). These two approaches can be most easily combined by using a fractional-step method. See `[claw/book/chap17/advdiff]` for an example.

In Section 19.5 we will also see that it is possible to solve a two-dimensional hyperbolic equation of the form

$$q_t + f(q)_x + g(q)_y = 0$$

by splitting it into two one-dimensional problems $q_t + f(q)_x = 0$ and $q_t + g(q)_y = 0$ and using one-dimensional high-resolution methods for each piece. The same idea extends to three space dimensions. In this context the fractional-step approach is called *dimensional splitting*. The theory developed in this chapter is also useful in analyzing these methods.

17.2 An Advection–Reaction Equation

To illustrate, we begin with a simple advection–reaction equation.

Example 17.1. Consider the linear equation

$$q_t + \bar{u} q_x = -\beta q, \tag{17.7}$$

with data $q(x, 0) = \overset{\circ}{q}(x)$. This would model, for example, the transport of a radioactive material in a fluid flowing at constant speed $\bar{u}$ down a pipe. The material decays as it flows along, at rate β. We can easily compute the exact solution of (17.7), since along the characteristic $dx/dt = \bar{u}$ we have $dq/dt = -\beta q$, and hence

$$q(x, t) = e^{-\beta t}\, \overset{\circ}{q}(x - \bar{u}t). \tag{17.8}$$

17.2.1 An Unsplit Method

Before discussing fractional-step methods in more detail, we first present an *unsplit* method for (17.7), which more clearly models the correct equation. An obvious extension of the upwind method for advection would be (assuming $\bar{u} > 0$)

$$Q_i^{n+1} = Q_i^n - \frac{\bar{u}\,\Delta t}{\Delta x}\left(Q_i^n - Q_{i-1}^n\right) - \Delta t\, \beta Q_i^n. \tag{17.9}$$

This method is first-order accurate and stable for $0 < \bar{u}\,\Delta t/\Delta x \le 1$; see Exercise 8.3.

A second-order Lax–Wendroff-style method can be developed by using the Taylor series

$$q(x, t + \Delta t) \approx q(x, t) + \Delta t\, q_t(x, t) + \frac{1}{2}\Delta t^2 q_{tt}(x, t). \tag{17.10}$$

As in the derivation of the Lax–Wendroff method in Section 6.1, we must compute q_{tt} from the PDE. Differentiating q_t gives

$$q_{tt} = -\bar{u} q_{xt} - \beta q_t, \qquad q_{tx} = -\bar{u} q_{xx} - \beta q_x,$$

and combining these, we obtain

$$q_{tt} = \bar{u}^2 q_{xx} + 2\bar{u}\beta q_x + \beta^2 q. \tag{17.11}$$

Note that this is more easily obtained by using

$$\partial_t q = (-\bar{u}\partial_x - \beta)q,$$

and hence

$$\partial_t^2 q = (-\bar{u}\partial_x - \beta)^2 q = \big(\bar{u}^2\partial_x^2 + 2\bar{u}\beta\partial_x + \beta^2\big)q. \tag{17.12}$$

Using this expression for q_{tt} in (17.10) gives

$$\begin{aligned} q(x, t + \Delta t) &\approx q - \Delta t\,(\bar{u}q_x + \beta q) + \frac{1}{2}\Delta t^2(\bar{u}^2 q_{xx} + 2\bar{u}\beta q_x + \beta^2 q) \\ &= \left(1 - \Delta t\,\beta + \frac{1}{2}\Delta t^2\beta^2\right) q - \Delta t\,\bar{u}\,(1 - \Delta t\,\beta)\,q_x + \frac{1}{2}\Delta t^2\,\bar{u}^2 q_{xx}. \end{aligned} \tag{17.13}$$

We can now approximate x-derivatives by finite differences to obtain the *second-order method*

$$\begin{aligned} Q_i^{n+1} = {}&\left(1 - \Delta t\,\beta + \frac{1}{2}\Delta t^2\beta^2\right) Q_i^n - \frac{\bar{u}\,\Delta t}{2\,\Delta x}\,(1 - \Delta t\,\beta)\big(Q_{i+1}^n - Q_{i-1}^n\big) \\ &+ \frac{\bar{u}^2\Delta t^2}{2\,\Delta x^2}\big(Q_{i-1}^n - 2Q_i^n + Q_{i+1}^n\big). \end{aligned} \tag{17.14}$$

In order to model the equation (17.7) correctly to second-order accuracy, we must properly model the interaction between the $\bar{u}q_x$ and the βq terms, which brings in the mixed term $\frac{1}{2}\Delta t^2\,\bar{u}\beta q_x$ in the Taylor series expansion.

For future use we also note that for (17.7) the full Taylor series expansion can be written as

$$q(x, t + \Delta t) = \sum_{j=0}^{\infty} \frac{(\Delta t)^j}{j!}\partial_t^j q(x, t) = \sum_{j=0}^{\infty} \frac{(\Delta t)^j}{j!}(-\bar{u}\partial_x - \beta)^j q(x, t), \tag{17.15}$$

which can be written formally as

$$q(x, t + \Delta t) = e^{-\Delta t\,(\bar{u}\partial_x + \beta)} q(x, t). \tag{17.16}$$

The operator $e^{-\Delta t\,(\bar{u}\partial_x + \beta)}$, which is defined via the Taylor series in (17.15), is called the *solution operator* for the equation (17.7) over a time step of length Δt. Applying this operator to any function of x gives the evolution of this data after time Δt has elapsed.

Note that the second derivative q_{tt} might be harder to compute for a more complicated problem, making it harder to develop second-order numerical methods. It is also not clear how to introduce limiters effectively into the unsplit method (17.14), as might be desirable in solving problems with discontinuous solutions. In some situations, it is possible to use the ideas of high-resolution methods based on Riemann solvers and also include the effects of source terms in the process of solving the Riemann problem. One approach of this form is presented in Section 17.14, and some others can be found in [34], [127], [128], [151], [155], [161], [162], [167], [168], [482], [216], [217], [284], [325], [381], [471]. In many cases, however, the simple fractional-step approach can often be effectively used, as described in the next section.

17.2.2 A Fractional-Step Method

A fractional-step method for (17.7) is applied by first splitting the equation into two *subproblems* that can be solved independently. For the advection–reaction problem (17.7) we would take these to be:

$$\text{Problem A:}\quad q_t + \bar{u} q_x = 0, \tag{17.17}$$

$$\text{Problem B:}\quad q_t = -\beta q. \tag{17.18}$$

The idea of the fractional-step method is to combine these by applying the two methods in an alternating manner. For more complicated problems this has great advantage over attempting to derive an unsplit method. If we split the general problem $q_t + f(q)_x = \psi(q)$ into the homogeneous conservation law and a simple ODE, then we can use standard methods for each. In particular, the high-resolution shock-capturing methods already developed can be used directly for the homogeneous conservation law, whereas trying to derive an unsplit method based on the same ideas while incorporating the source term directly can be more difficult.

As a simple example of the fractional-step procedure, suppose we use the upwind method for the A-step and the forward Euler for the ODE in the B-step for the advection–reaction problem. Then the simplest fractional-step method over one time step would consist of the following two stages:

$$\text{A-step:}\qquad Q_i^* = Q_i^n - \frac{\bar{u}\,\Delta t}{\Delta x}\left(Q_i^n - Q_{i-1}^n\right), \tag{17.19}$$

$$\text{B-step:}\quad Q_i^{n+1} = Q_i^* - \beta\,\Delta t\, Q_i^*. \tag{17.20}$$

Note that we first take a time step of length Δt with upwind, starting with initial data Q_i^n to obtain the intermediate value Q_i^*. Then we take a time step of length Δt using the forward Euler method, starting with the data Q^* obtained from the first stage.

It may seem that we have advanced the solution by time $2\,\Delta t$ after taking these two steps of length Δt. However, in each stage we used only some of the terms in the original PDE, and the two stages combined give a consistent approximation to solving the original equation (17.7) over a single time step of length Δt. To check this consistency, we can combine the two stages by eliminating Q^* to obtain a method in a more familiar form:

$$\begin{aligned} Q_i^{n+1} &= (1-\beta\,\Delta t)Q_i^* \\ &= (1-\beta\,\Delta t)\left[Q_i^n - \frac{\bar{u}\,\Delta t}{\Delta x}\left(Q_i^n - Q_{i-1}^n\right)\right] \\ &= Q_i^n - \frac{\bar{u}\,\Delta t}{\Delta x}\left(Q_i^n - Q_{i-1}^n\right) - \beta\,\Delta t\, Q_i^n + \frac{\bar{u}\beta\,\Delta t^2}{\Delta x}\left(Q_i^n - Q_{i-1}^n\right). \end{aligned} \tag{17.21}$$

The first three terms on the right-hand side agree with the unsplit method (17.9). The final term is $\mathcal{O}(\Delta t^2)$ (since $(Q_i^n - Q_{i-1}^n)/\Delta x \approx q_x = \mathcal{O}(1)$), and so a local truncation error analysis will show that this method, though slightly different from (17.9), is also consistent and first-order accurate on the original equation (17.7).

A natural question is whether we could improve the accuracy by using a more accurate method in each step. For example, suppose we use the Lax–Wendroff method in the A-step and the trapezoidal method, or the two-stage Runge–Kutta method, in the B-step. Would we then obtain a second-order accurate method for the original equation? For this particular equation, the answer is yes. In fact if we use pth-order accurate methods for each step, the result will be a pth-order accurate method for the full original equation. But this equation is very special in this regard, and this claim should seem surprising. One would think that splitting the equation into pieces in this manner would introduce some error that depends on the size of the time step Δt and is independent of how well we then approximate the subproblem in each step. In general this is true – there is a "splitting error" that in general would be $\mathcal{O}(\Delta t)$ for the type of splitting used above, and so the resulting fractional-step method will be only first-order accurate, no matter how well we then approximate each step. This will be analyzed in more detail below.

For the case of equation (17.7) there is no splitting error. This follows from the observation that we can solve (17.7) over any time period Δt by first solving the equation (17.17) over time Δt, and then using the result as data to solve the equation (17.18) over time Δt. To verify this, let $u^*(x, \Delta t)$ be the exact solution to the A-problem,

$$\begin{aligned} q^*_t + \bar{u} q^*_x &= 0, \\ q^*(x,0) &= \overset{\circ}{q}(x). \end{aligned} \tag{17.22}$$

We use a different symbol $q^*(x,t)$ for the solution to this problem rather than $q(x,t)$, which we reserve for the exact solution to the original problem.

Then we have

$$q^*(x, \Delta t) = \overset{\circ}{q}(x - \bar{u}\,\Delta t).$$

If we now use this as data in solving the B-problem (17.18), we will be solving a different equation,

$$q^{**}_t = -\beta q^{**} \tag{17.23}$$

with initial data

$$q^{**}(x,0) = q^*(x, \Delta t) = \overset{\circ}{q}(x - \bar{u}\,\Delta t).$$

This is just an ODE at each point x, and the solution is

$$q^{**}(x, \Delta t) = e^{-\beta\,\Delta t}\,\overset{\circ}{q}(x - \bar{u}\,\Delta t).$$

Comparing this with (17.8), we see that we have indeed recovered the solution to the original problem by this two-stage procedure.

Physically we can interpret this as follows. Think of the original equation as modeling a radioactive tracer that is advecting with constant speed $\bar{u}$ (carried along in a fluid, say) and also decaying with rate β. Since the decay properties are independent of the position x, we

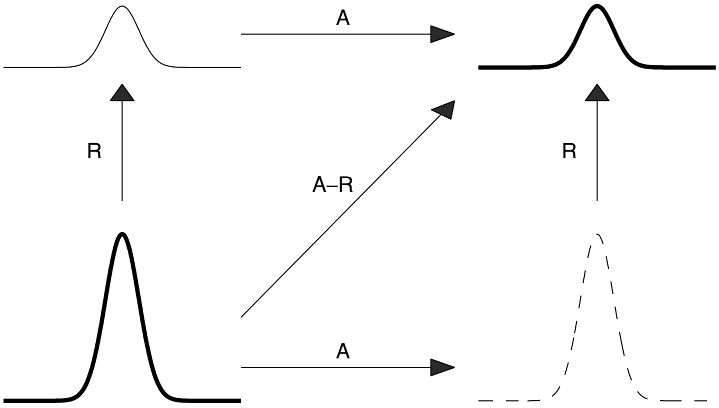


Fig. 17.1. Illustration of a fractional-step procedure for the advection–reaction equation (17.7) when there is no splitting error. The pulse shown in the lower left simultaneously advects and decays as indicated by the diagonal arrow labeled A-R. The same result is obtained if the pulse first advects to the right following the arrow A and then is allowed to decay via the reaction term, following the arrow R, or if the reaction and advection are performed in the opposite order.

can think of first advecting the tracer over time Δt without allowing any decay, and then holding the fluid and tracer stationary while we allow it to decay for time Δt. We will get the same result, and this is what we have done in the fractional-step method. Figure 17.1 illustrates this.

We would also get the same result if we first allowed the tracer to decay at the initial location and then advected the decayed profile, which amounts to switching the order in which the two subproblems are solved (see Figure 17.1). We say that the solution operators for the two subproblems *commute*, since we can apply them in either order and get the same result. In general if these operators commute, then there is no splitting error, a fact that we will investigate more formally in Section 17.3. (Here we are only discussing the Cauchy problem. Boundary conditions can further complicate the situation; see Section 17.9).

Another way to examine the splitting error, which must be used more generally when we do not know the exact solution to the equations involved, is to use Taylor series expansions. (This approach can be used also for nonlinear problems.) If we look at a time step of length Δt, then solving the A-equation gives

$$\begin{aligned} q^*(x,\Delta t) &= q^*(x,0) + \Delta t\, q^*_t(x,0) + \frac{1}{2}\Delta t^2 q^*_{tt}(x,0) + \cdots \\ &= q^*(x,0) - \bar{u}\,\Delta t\, q^*_x(x,0) + \frac{1}{2}\bar{u}^2\Delta t^2 q^*_{xx}(x,0) - \cdots \\ &= q(x,0) - \bar{u}\,\Delta t\, q_x(x,0) + \frac{1}{2}\bar{u}^2\Delta t^2 q_{xx}(x,0) - \cdots . \end{aligned} \tag{17.24}$$

Similarly, if we solve the ODE problem (17.23) with general initial data $q^{**}(x,0)$, we obtain

$$\begin{aligned} q^{**}(x,\Delta t) &= q^{**}(x,0) + \Delta t\, q^{**}_t(x,0) + \frac{1}{2}\Delta t^2 q^{**}_{tt}(x,0) + \cdots \\ &= \left(1 - \beta\,\Delta t + \frac{1}{2}\beta^2\Delta t^2 + \cdots\right) q^{**}(x,0). \end{aligned} \tag{17.25}$$

If we now use the result from (17.24) as the initial data in (17.25), we obtain

$$\begin{aligned} q^{**}(x,\Delta t) &= \left(1 - \Delta t\,\beta + \frac{1}{2}\Delta t^2\beta^2 - \cdots\right)\Big(q(x,0) - \bar{u}\,\Delta t\,q_x(x,0) \\ &\qquad + \frac{1}{2}\bar{u}^2\Delta t^2 q_{xx}(x,0) + \cdots\Big) \\ &= q - \Delta t\,(\bar{u}q_x + \beta q) + \frac{1}{2}\Delta t^2(\bar{u}^2 q_{xx} + 2\bar{u}\beta q_x + \beta^2 q) + \cdots. \end{aligned} \tag{17.26}$$

Comparing this with the Taylor series expansion (17.13) that we used in deriving the unsplit Lax–Wendroff method shows that this agrees with $q(x, \Delta t)$, at least for the three terms shown, and in fact to all orders.

Note that the mixed term $\bar{u}\beta\,\Delta t^2 q_x$ needed in the q_{tt}-term from (17.11) now arises naturally from taking the product of the two Taylor series (17.24) and (17.25). In fact, we see that for this simple equation we can write (17.25) as

$$q^{**}(x,\Delta t) = e^{-\beta\,\Delta t}\,q^{**}(x,0),$$

while (17.24) can be written formally as

$$q^{*}(x,\Delta t) = e^{-\bar{u}\,\Delta t\,\partial_x}\overset{\circ}{q}(x).$$

If we now use $q^*(x, \Delta t)$ as the data $q^{**}(x, 0)$ as we do in the fractional-step method, we obtain

$$q^{**}(x,\Delta t) = e^{-\beta\,\Delta t}\,e^{-\bar{u}\,\Delta t\,\partial_x}\overset{\circ}{q}(x).$$

Multiplying out the Taylor series as we did in (17.26) verifies that these exponentials satisfy the usual rule, so that to compute the product we need only add the exponents, i.e.,

$$q^{**}(x,\Delta t) = e^{-\Delta t\,(\bar{u}\partial_x + \beta)}\overset{\circ}{q}(x).$$

The exponential appearing here is exactly the solution operator for the original equation, as in (17.16), and so again we see that $q^{**}(x, \Delta t) = q(x, \Delta t)$ and there is no splitting error.

The fact that there is no splitting error for the problem (17.7) is a reflection of the fact that, for this problem, the solution operator for the full problem is exactly equal to the product of the solution operators of the two subproblems (17.17) and (17.18). This is not generally true for other problems.

Example 17.2. Suppose we modify the equation slightly so that the decay rate β depends on x,

$$q_t + \bar{u}q_x = -\beta(x)q. \tag{17.27}$$

Then our previous argument for the lack of a splitting error breaks down – advecting the tracer a distance $\bar{u}\,\Delta t$ and then allowing it to decay, with rates given by the values of β

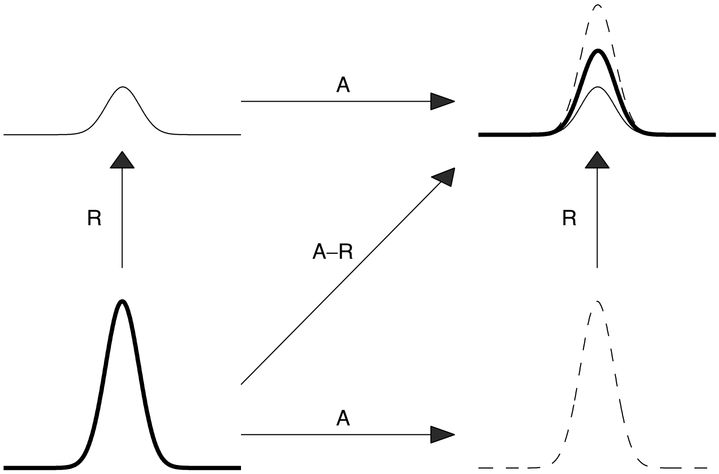


Fig. 17.2. Illustration of a fractional-step procedure for the advection–reaction equation (17.27), where there is a splitting error because the decay rate β depends on x. The pulse shown in the lower left simultaneously advects and decays, as indicated by the diagonal arrow labeled A-R. Different results are obtained if the pulse first advects to the right following the arrow A and then is allowed to decay via the reaction term evaluated at the final position, following the arrow R, or if the reaction and advection are performed in the opposite order.

at the final positions, will not in general give the same result as when the decay occurs continuously as it advects, using the instantaneous rate given by $\beta(x)$ at each point passed.

Figure 17.2 illustrates the fact that solution operators for the two subproblems do not commute, shown for the case $\beta(x) = 1 - x$ over $0 \le x \le 1$, so that the decay rate is smaller for larger x. First advecting and then reacting gives too little decay, while first reacting and then advecting gives too much decay. Note that this is shown for very large Δt in Figure 17.2 in order to illustrate the effect clearly. With a numerical fractional-step method we would be using much smaller time steps to solve the problem over this time period, in each step advecting by a small amount and then reacting, so that reasonable results could still be obtained, though formally only first-order accurate as the time step is reduced. See Section 17.5 for some numerical results.

The accuracy of the fractional-step method on (17.27) can be analyzed formally using Taylor series expansions again. Rather than developing this expansion for this particular example, we will first examine the more general case and then apply it to that case.

17.3 General Formulation of Fractional-Step Methods for Linear Problems

Consider a more general linear PDE of the form

$$q_t = (\mathcal{A} + \mathcal{B})q, \tag{17.28}$$

where $\mathcal{A}$ and $\mathcal{B}$ may be differential operators, e.g., $\mathcal{A} = -\bar{u}\partial_x$ and $\mathcal{B} = -\beta(x)$ in Example 17.2. For simplicity suppose that $\mathcal{A}$ and $\mathcal{B}$ do not depend explicitly on t, e.g., $\beta(x)$ is a function of x but not of t. Then we can compute that

$$q_{tt} = (\mathcal{A} + \mathcal{B})q_t = (\mathcal{A} + \mathcal{B})^2 q,$$

and in general

$$\partial_t^j q = (\mathcal{A}+\mathcal{B})^j q. \tag{17.29}$$

We have used this idea before in calculating Taylor series, e.g., in (17.12).

Note that if $\mathcal{A}$ or $\mathcal{B}$ do depend on t, then we would have to use the product rule, e.g.,

$$q_{tt} = (\mathcal{A}+\mathcal{B})q_t + (\mathcal{A}_t+\mathcal{B}_t)q,$$

and everything would become more complicated. Also note that if the problem is nonlinear then the Taylor series expansion can still be used if the solution is smooth, but we don't generally have a simple relation of the form (17.29).

In our simple case we can write the solution at time t using Taylor series as

$$\begin{aligned} q(x,\Delta t) &= q(x,0) + \Delta t(\mathcal{A}+\mathcal{B})q(x,0) + \frac{1}{2}\Delta t^2(\mathcal{A}+\mathcal{B})^2 q(x,0) + \cdots \\ &= \left(I + \Delta t\,(\mathcal{A}+\mathcal{B}) + \frac{1}{2}\Delta t^2(\mathcal{A}+\mathcal{B})^2 + \cdots\right) q(x,0) \\ &= \sum_{j=0}^{\infty} \frac{\Delta t^j}{j!}(\mathcal{A}+\mathcal{B})^j q(x,0), \end{aligned} \tag{17.30}$$

which formally can be written as

$$q(x,\Delta t) = e^{\Delta t\,(\mathcal{A}+\mathcal{B})} q(x,0).$$

With the fractional-step method, we instead compute

$$q^*(x,\Delta t) = e^{\Delta t\,\mathcal{A}}\, q(x,0),$$

and then

$$q^{**}(x,\Delta t) = e^{\Delta t\,\mathcal{B}} q^*(x,\Delta t) = e^{\Delta t\,\mathcal{B}} e^{\Delta t\,\mathcal{A}}\, q(x,0),$$

and so the *splitting error* is

$$q(x,\Delta t) - q^{**}(x,\Delta t) = \left(e^{\Delta t\,(\mathcal{A}+\mathcal{B})} - e^{\Delta t\,\mathcal{B}} e^{\Delta t\,\mathcal{A}}\right) q(x,0). \tag{17.31}$$

This should be calculated using the Taylor series expansions. We have (17.30) already, while

$$\begin{aligned} q^{**}(x,\Delta t) &= \left(I + \Delta t\,\mathcal{B} + \frac{1}{2}\Delta t^2\mathcal{B}^2 + \cdots\right)\left(I + \Delta t\,\mathcal{A} + \frac{1}{2}\Delta t^2\mathcal{A}^2 + \cdots\right) q(x,0) \\ &= \left(I + \Delta t\,(\mathcal{A}+\mathcal{B}) + \frac{1}{2}\Delta t^2(\mathcal{A}^2 + 2\mathcal{B}\mathcal{A} + \mathcal{B}^2) + \cdots\right) q(x,0). \end{aligned} \tag{17.32}$$

The $I + \Delta t\,(\mathcal{A}+\mathcal{B})$ terms agree with (17.30). In the Δt^2 term, however, the term from (17.30) is

$$\begin{aligned} (\mathcal{A}+\mathcal{B})^2 &= (\mathcal{A}+\mathcal{B})(\mathcal{A}+\mathcal{B}) \\ &= \mathcal{A}^2 + \mathcal{A}\mathcal{B} + \mathcal{B}\mathcal{A} + \mathcal{B}^2. \end{aligned} \tag{17.33}$$

In general this is *not* the same as

$$\mathcal{A}^2 + 2\mathcal{B}\mathcal{A} + \mathcal{B}^2,$$

and so the splitting error is

$$q(x, \Delta t) - q^{**}(x, \Delta t) = \frac{1}{2}\Delta t^2(\mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A})q(x, 0) + \mathcal{O}(\Delta t^3). \qquad (17.34)$$

The splitting error depends on the *commutator* $\mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A}$ and is zero only in the special case when the differential operators $\mathcal{A}$ and $\mathcal{B}$ commute (in which case it turns out that all the higher-order terms in the splitting error also vanish).

Example 17.3. For the problem considered in Example 17.1,

$$\mathcal{A} = -\bar{u}\partial_x \quad \text{and} \quad \mathcal{B} = -\beta.$$

We then have $\mathcal{A}\mathcal{B}q = \mathcal{B}\mathcal{A}q = \bar{u}\beta q_x$. These operators commute for β constant, and there is no splitting error.

Example 17.4. Now suppose $\beta = \beta(x)$ depends on x as in Example 17.2. Then we have

$$\mathcal{A}\mathcal{B}q = \bar{u}\partial_x(\beta(x)q) = \bar{u}\beta(x)q_x + \bar{u}\beta'(x)q,$$

while

$$\mathcal{B}\mathcal{A}q = \beta(x)\bar{u}q_x.$$

These are not the same unless $\beta'(x) = 0$. In general the splitting error will be

$$q(x, \Delta t) - q^{**}(x, \Delta t) = \frac{1}{2}\Delta t^2\bar{u}\beta'(x)q(x, 0) + \mathcal{O}(\Delta t^3).$$

If we now design a fractional-step method based on this splitting, we will see that the splitting error alone will introduce an $\mathcal{O}(\Delta t^2)$ error in each time step, which can be expected to accumulate to an $\mathcal{O}(\Delta t)$ error after the $T/\Delta t$ time steps needed to reach some fixed time T (in the best case, assuming the method is stable). Hence even if we solve each subproblem *exactly* within the fractional-step method, the resulting method will be only first-order accurate. If the subproblems are actually solved with numerical methods that are sth-order accurate, the solution will still only be first-order accurate no matter how large s is. At least this is true asymptotically as the mesh spacing tends to zero. In practice results that are essentially second-order accurate are observed, for reasons described in Section 17.5.

Of course this order of accuracy can only be obtained for smooth solutions. We are often interested in problems where the solution is not smooth, in which case a lower order of accuracy is generally observed. In this case the ability to easily use high-resolution methods for the hyperbolic portion of the problem is an advantage of the fractional-step approach. On the other hand, it is not so clear that the method even converges in this case, since the above arguments based on Taylor series expansions do not apply directly. For some convergence results, see for example [258], [442], [443].

17.4 Strang Splitting

The above form of fractional-step method, sometimes called the *Godunov splitting*, in general is only first-order accurate formally. It turns out that a slight modification of the splitting idea will yield second-order accuracy quite generally (assuming each subproblem is solved with a method of at least this accuracy). The idea is to solve the first subproblem $q_t = \mathcal{A}q$ over only a half time step of length $\Delta t/2$. Then we use the result as data for a full time step on the second subproblem $q_t = \mathcal{B}q$, and finally take another half time step on $q_t = \mathcal{A}q$. We can equally well reverse the roles of $\mathcal{A}$ and $\mathcal{B}$ here. This approach is often called *Strang splitting*, as it was popularized in a paper by Strang [426] on solving multidimensional problems.

To analyze the Strang splitting, note that we are now approximating the solution operator $e^{\Delta t\,(\mathcal{A}+\mathcal{B})}$ by $e^{\frac{1}{2}\Delta t\,\mathcal{A}}e^{\Delta t\,\mathcal{B}}e^{\frac{1}{2}\Delta t\,\mathcal{A}}$. Taylor series expansion of this product shows that

$$\begin{aligned} e^{\frac{1}{2}\Delta t\,\mathcal{A}}e^{\Delta t\,\mathcal{B}}e^{\frac{1}{2}\Delta t\,\mathcal{A}} &= \left(I + \frac{1}{2}\Delta t\,\mathcal{A} + \frac{1}{8}\Delta t^2\mathcal{A}^2 + \cdots\right)\left(I + \Delta t\,\mathcal{B} + \frac{1}{2}\Delta t^2\mathcal{B}^2 + \cdots\right) \\ &\quad \times \left(I + \frac{1}{2}\Delta t\,\mathcal{A} + \frac{1}{8}\Delta t^2\mathcal{A}^2 + \cdots\right) \\ &= I + \Delta t\,(\mathcal{A}+\mathcal{B}) + \frac{1}{2}\Delta t^2(\mathcal{A}^2 + \mathcal{A}\mathcal{B} + \mathcal{B}\mathcal{A} + \mathcal{B}^2) + \mathcal{O}(\Delta t^3). \end{aligned} \tag{17.35}$$

Comparing with (17.30), we see that the $\mathcal{O}(\Delta t^2)$ term is now captured correctly. The $\mathcal{O}(\Delta t^3)$ term is not correct in general, however, unless $\mathcal{A}\mathcal{B} = \mathcal{B}\mathcal{A}$.

Note that over several time steps we can simplify the expression obtained with the Strang splitting. After n steps we have

$$Q^n = \left(e^{\frac{1}{2}\Delta t\,\mathcal{A}}e^{\Delta t\,\mathcal{B}}e^{\frac{1}{2}\Delta t\,\mathcal{A}}\right)\left(e^{\frac{1}{2}\Delta t\,\mathcal{A}}e^{\Delta t\,\mathcal{B}}e^{\frac{1}{2}\Delta t\,\mathcal{A}}\right)\cdots\left(e^{\frac{1}{2}\Delta t\,\mathcal{A}}e^{\Delta t\,\mathcal{B}}e^{\frac{1}{2}\Delta t\,\mathcal{A}}\right)Q^0 \tag{17.36}$$

repeated n times. Dropping the parentheses and noting that $e^{\frac{1}{2}\Delta t\,\mathcal{A}}e^{\frac{1}{2}\Delta t\,\mathcal{A}} = e^{\Delta t\,\mathcal{A}}$, we obtain

$$Q^n = e^{\frac{1}{2}\Delta t\,\mathcal{A}}\,e^{\Delta t\,\mathcal{B}}e^{\Delta t\,\mathcal{A}}e^{\Delta t\,\mathcal{B}}e^{\Delta t\,\mathcal{A}}\cdots e^{\Delta t\,\mathcal{B}}\,e^{\frac{1}{2}k\mathcal{A}}Q^0. \tag{17.37}$$

This differs from the Godunov splitting only in the fact that we start and end with a half time step on $\mathcal{A}$, rather than starting with a full step on $\mathcal{A}$ and ending with $\mathcal{B}$.

Another way to achieve this same effect is to simply take steps of length Δt on each problem, as in the first-order splitting, but to alternate the order of these steps in alternate time steps, e.g.,

$$\begin{aligned} Q^1 &= e^{\Delta t\,\mathcal{B}}e^{\Delta t\,\mathcal{A}}Q^0, \\ Q^2 &= e^{\Delta t\,\mathcal{A}}e^{\Delta t\,\mathcal{B}}Q^1, \\ Q^3 &= e^{\Delta t\,\mathcal{B}}e^{\Delta t\,\mathcal{A}}Q^2, \\ Q^4 &= e^{\Delta t\,\mathcal{A}}e^{\Delta t\,\mathcal{B}}Q^3, \\ &\;\;\vdots \end{aligned}$$

If we take an even number of time steps, then we obtain

$$\begin{aligned}Q^n &= (e^{\Delta t\,\mathcal{A}}e^{\Delta t\,\mathcal{B}})(e^{\Delta t\,\mathcal{B}}e^{\Delta t\,\mathcal{A}})(e^{\Delta t\,\mathcal{A}}e^{\Delta t\,\mathcal{B}})(e^{\Delta t\,\mathcal{B}}e^{\Delta t\,\mathcal{A}})\cdots(e^{\Delta t\,\mathcal{A}}e^{\Delta t\,\mathcal{B}})(e^{\Delta t\,\mathcal{B}}e^{\Delta t\,\mathcal{A}})Q^0\\ &= e^{\Delta t\,\mathcal{A}}(e^{\Delta t\,\mathcal{B}}e^{\Delta t\,\mathcal{B}})(e^{\Delta t\,\mathcal{A}}e^{\Delta t\,\mathcal{A}})(e^{\Delta t\,\mathcal{B}}e^{\Delta t\,\mathcal{B}})\cdots\big(e^{\Delta t\,\mathcal{B}}e^{\Delta t\,\mathcal{B}}\big)e^{\Delta t\,\mathcal{A}}Q^0.\end{aligned}$$

Since $e^{\Delta t\,\mathcal{B}}e^{\Delta t\,\mathcal{B}} = e^{2\,\Delta t\,\mathcal{B}}$, this is essentially the same as (17.36) but with $\frac{1}{2}\Delta t$ replaced by Δt. This is generally more efficient than the approach of (17.36), since a single step with the numerical method approximating $e^{\Delta t\,\mathcal{A}}$ is typically cheaper than two steps of length $\Delta t/2$. On the other hand, this form is more difficult to implement with variable time steps Δt, as are often used in practice. An even number of steps must be taken and the value of Δt in each pair of steps must be the same, in order to obtain the desired cancellation of errors.

In CLAWPACK, either the Godunov splitting or the Strang splitting (implemented in the form (17.36)) can be selected by setting `method(5) = 1` or `2` respectively. In this case a subroutine `src1.f` must be provided that solves the $q_t = \psi(q)$ subproblem arising from the source terms.

17.5 Accuracy of Godunov and Strang Splittings

The fact that the Strang splitting is so similar to the first-order splitting suggests that the first-order splitting is not really so bad, and in fact it is not. While formally only first-order accurate, the coefficient of the $\mathcal{O}(\Delta t)$ term may be much smaller than coefficients in the second-order terms arising from discretization of $e^{\Delta t\,\mathcal{A}}$ and $e^{\Delta t\,\mathcal{B}}$.

For this reason the simpler and more efficient Godunov splitting is often sufficient. It is also easier to implement boundary conditions properly with the Godunov splitting, as discussed in Section 17.9.

Example 17.5. Figure 17.3 shows results at time $t = 0.5$ from solving the problem (17.27) with $\bar{u} = 1$, $\beta(x) = 1 - x$, and initial data consisting of a Gaussian pulse centered at $x = 0.25$. The Godunov and Strang splittings are compared, where in each case the

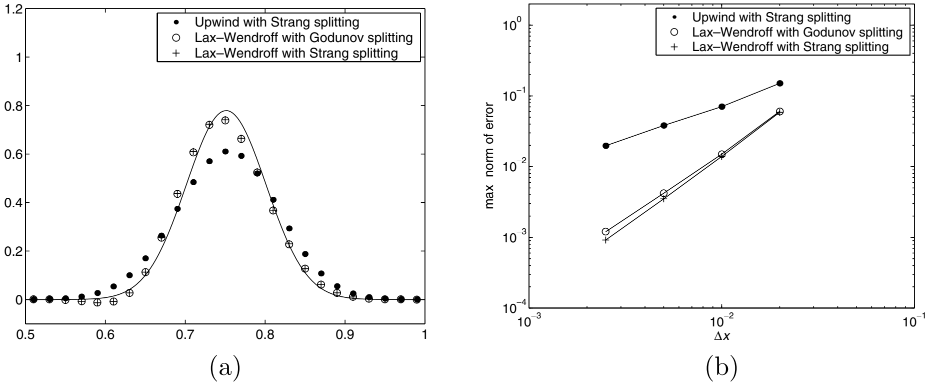


Fig. 17.3. Comparison of results with three methods applied to the problem (17.27). (a) Computed and true solution for $\Delta x = 0.02$. (b) Log–log plot of max-norm errors vs. Δx. Note that the Godunov splitting is essentially as accurate as the Strang splitting for this problem. `[claw/book/chap17/nocommute]`

Lax–Wendroff method is used for the advection equation and the second-order two-stage Runge–Kutta method is used for the source term. On this grid with $\Delta x = 0.02$, the results are visually indistinguishable, even though the Godunov splitting is formally only first-order accurate.

For contrast, Figure 17.3(a) also shows the results obtained if the first-order upwind method is used in place of the Lax–Wendroff method in the Strang splitting. This first-order method causes a substantial smearing of the solution.

Figure 17.3(b) shows log–log plots of the error in the max norm for each of these three methods as the grid is refined. For very fine grids the Godunov splitting is slightly less accurate and asymptotically approaches first-order, but even for the finest grid used ($\Delta x = 1/400$) it is only slightly less accurate than the Strang splitting. By contrast the upwind method is much less accurate. As in the discussion of the accuracy of limiters in Section 8.5, it is important to realize that order of accuracy is not the full story.

17.6 Choice of ODE Solver

Consider the Godunov splitting, and suppose we have already obtained Q^* from Q^n by solving the conservation law (17.5). We now wish to advance Q_i^* to Q_i^{n+1} by solving the ODE $q_t = \psi(q)$ over time Δt in each grid cell. In some cases this equation can be solved exactly. For the system (17.4), for example, the source terms alone yield

$$\begin{aligned} p_t &= -\frac{2K_0 u}{r}, \\ u_t &= 0. \end{aligned} \tag{17.38}$$

Since u is constant in this system, the value of p_t is constant and this ODE is easily solved exactly, yielding

$$\begin{aligned} p_i^{n+1} &= p_i^* - \Delta t\,(2K_0 u_i^*)/r_i, \\ u_i^{n+1} &= u_i^*. \end{aligned} \tag{17.39}$$

For more complicated source terms, e.g., those arising from chemical kinetics with many interacting species, it will be necessary to use a numerical ODE solver in each grid cell. We typically want to use a method that is at least second-order accurate to maintain overall accuracy. A wide variety of ODE solvers are available for systems of the general form $y' = \psi(y)$ where $y(t) \in \mathbb{R}^m$. Note, however, that in general we cannot use multistep methods that require more than one level of data (e.g., y^{n-1} as well as y^n) to generate the solution y^{n+1} at the next time level. This is because we only have data Q_i^* to use in computing Q_i^{n+1}. Previous values (e.g., Q_i^n or Q_i^* from the previous time step) are not suitable to use in the context of multistep methods, because Q_i^* is computed from Q_i^n by solving a different equation (the conservation law (17.5)) than the ODE we are now attempting to approximate.

In many cases a simple explicit Runge–Kutta method can be effectively used. These are multistage one-step methods that generate intermediate values as needed to construct higher-order approximations. A simple second-order accurate two-stage method is often sufficient for use with high-resolution methods, for example the classical method

$$\begin{aligned} Q_i^{**} &= Q_i^* + \frac{\Delta t}{2}\psi(Q_i^*), \\ Q_i^{n+1} &= Q_i^* + \Delta t\, \psi(Q_i^{**}). \end{aligned} \tag{17.40}$$

One must ensure that the explicit method is stable with the time step Δt being used, or perhaps take N time steps of (17.40) using a smaller step size $\Delta t/N$ to advance Q_i^* to Q_i^{n+1} stably.

17.7 Implicit Methods, Viscous Terms, and Higher-Order Derivatives

If the ODEs $q_t = \psi(q)$ are *stiff*, as discussed in Section 17.10, then it may be necessary to use an implicit method in this step in order to use a reasonable time step. In this case other numerical issues arise, and even a stable implicit method may give poor results, as illustrated in Section 17.16.

A natural implicit method to consider is the trapezoidal method, a second-order accurate one-step method that takes the form

$$Q_i^{n+1} = Q_i^* + \frac{\Delta t}{2}\left[\psi(Q_i^*) + \psi\left(Q_i^{n+1}\right)\right]. \tag{17.41}$$

Note, by the way, an advantage of the fractional-step approach for stiff equations. While this is an implicit method, the equations obtained in the ith cell are decoupled from the equations in every other cell, and so these equations can be solved relatively easily. The coupling between grid cells arises only in the hyperbolic part of the equation, which can still be solved with an explicit high-resolution method.

In some cases the source term ψ may depend on derivatives of q as well as the pointwise value, for example if we wish to solve a viscous equation $q_t + f(q)_x = \mu q_{xx}$ by a fractional-step approach. In this case the derivatives will have to be discretized, bringing in values of Q at neighboring grid cells. For example, the term $\psi(Q_i^{n+1})$ in (17.41) would be replaced by

$$\mu\left(Q_{i-1}^{n+1} - 2Q_i^{n+1} + Q_{i+1}^{n+1}\right)/\Delta x^2, \tag{17.42}$$

and similarly for $\psi(Q_i^*)$. The trapezoidal method (17.41) would then become the Crank–Nicolson method (4.13), and would require solving a tridiagonal linear system. It is generally necessary to use an implicit method for source terms involving second-order derivatives, since an explicit method would require $\Delta t = \mathcal{O}(\Delta x^2)$. With higher-order derivatives even smaller time steps would be required with an explicit method. The first-order hyperbolic part typically allows $\Delta t = \mathcal{O}(\Delta x)$, based on the CFL condition, and we generally hope to take time steps of this magnitude.

Although the trapezoidal method is second-order accurate and A-stable, it is only marginally stable in the stiff case, and this can lead to problems in the context of stiff hyperbolic equations, as illustrated in Section 17.16. For this reason the so-called *TR-BDF2 method* is generally recommended as a second-order implicit method. This is a two-stage Runge–Kutta method that combines one step of the trapezoidal method over time $\Delta t/2$ with a step of the second-order BDF method, using the intermediate result as another time level. For the ODE $q_t = \psi(q)$ this takes the form

$$\begin{aligned} Q_i^{**} &= Q_i^* + \frac{\Delta t}{2}[\psi(Q_i^*) + \psi(Q_i^{**})], \\ Q_i^{n+1} &= \frac{1}{3}\big[4Q_i^{**} - Q_i^* + \Delta t\, \psi\big(Q_i^{n+1}\big)\big]. \end{aligned} \tag{17.43}$$

If ψ represents viscous terms, say $\psi = \mu q_{xx}$, then again this will have to be discretized using terms of the form (17.42), leading to tridiagonal systems in each stage of the Runge–Kutta method.

An example illustrating the superiority of this method over the Crank–Nicolson method for handling a diffusion term is given in [462], for a reaction–diffusion–advection equation arising in a model of chemotaxis in bacterial growth. Another example of how the trapezoidal method can fail is given in Section 17.16.

17.8 Steady-State Solutions

There are some other potential pitfalls in using a fractional-step method to handle source terms. In this section we consider some of these in relation to computing a steady-state solution, one in which $q_t(x,t) \equiv 0$ and the function $q(x,t)$ is independent of time. For the homogeneous constant-coefficient linear hyperbolic equation $q_t + Aq_x = 0$, if $q_t = 0$ then $q_x = 0$ also and the only steady-state solutions are the constant functions. When a source term is added, there can be more interesting steady-state solutions. Consider the advection–reaction equation (17.7) from Section 17.2, $q_t + \bar{u}q_x = -\beta q$. Setting $q_t = 0$ gives the ODE $q_x = -(\beta/\bar{u})q$, and hence this has the steady-state solution

$$q(x,t) = Ce^{-(\beta/\bar{u})x}. \tag{17.44}$$

In practice we would have a finite domain and some boundary conditions that must also be satisfied. Consider the same PDE on the domain $0 < x < 1$ with initial data $\mathring{q}(x)$ and the boundary condition

$$q(0,t) = g_0(t)$$

at the inflow boundary. The general solution is

$$q(x,t) = \begin{cases} e^{-\beta t}\, \mathring{q}(x - \bar{u}t) & \text{if } t < x/\bar{u}, \\ e^{-(\beta/\bar{u})x}\, g_0(t - x/\bar{u}) & \text{if } t > x/\bar{u}. \end{cases} \tag{17.45}$$

In the special case $g_0(t) \equiv C$, some constant, the solution $q(x,t)$ will reach the steady state (17.44) for all $t > 1/\bar{u}$, regardless of the initial conditions.

Recall the physical interpretation of this equation as the advection at velocity $\bar{u}$ of a radiaoactive tracer that decays at rate β. The boundary condition $q(0,t) = C$ corresponds to inflowing fluid having concentration C at all times. Up to time $t = 1/\bar{u}$ the initial data also has an effect on the solution, but after this time all of the fluid (and tracer) initially

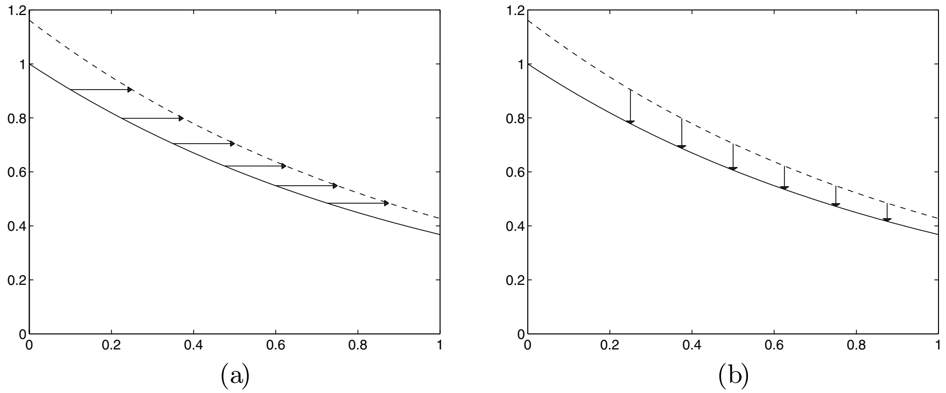


Fig. 17.4. Balance between advection and decay in a steady state solution. (a) The steady-state solution $q(x) = e^{-(\beta/\bar{u})x}$ is shown as the solid line. After advection by a distance $\bar{u}\,\Delta t$ to the right, $q^*(x) = q(x - \bar{u}\,\Delta t)$, the dashed line. (b) The advected solution decays by $e^{-\beta\Delta t}$, reproducing $q(x) = e^{-\beta\Delta t} q^*(x)$.

in the domain $0 < x < 1$ has flowed out the boundary at $x = 1$, and the steady state is reached.

It is important to note that being in a steady state does not in general mean that nothing is happening. In the above example, new tracer is constantly being introduced at $x = 0$, is advected downstream, and decays. The steady state results from a *balance* between the advection and decay processes. The terms $\bar{u}q_x$ and $-\beta q$ are both nonzero but cancel out. This balance is illustrated in Figure 17.4.

This suggests that we may have difficulties with a fractional-step method, where we first solve the advection equation ignoring the reactions and then solve the reaction equation ignoring the advection. Even if we start with the exact steady-state solution, each of these steps can be expected to make a change in the solution. In principle the two effects should exactly cancel out, but numerically they typically will not, since very different numerical techniques are used in each step.

In practice we generally don't know the steady-state solution, so we cannot use this as initial data. Instead we wish to determine the steady state by starting with some arbitrary initial data (perhaps some approximation to the steady state) and then marching forward in time until a steady state is reached. This can be viewed as an iterative method for solving the steady-state problem obtained by setting $q_t = 0$ in the equation, with each time step being one iteration. If all we care about is the steady-state solution, then this may not be a very efficient iterative method to use. A method designed specifically for the steady-state solution may be preferable. Such a method would be designed to converge as rapidly as possible to an accurate steady-state solution without necessarily giving an accurate time-dependent solution along the way. The study of such methods is a major topic in its own right and is not discussed further here.

However, time-marching methods are often used to compute steady-state solutions. For relatively small problems where computational efficiency is not an issue, it may be more cost-effective to use an existing time-marching code than to develop a new code specifically for the steady-state problem. One may also be interested in related time-dependent issues such as how convergence to steady state occurs in the physical system, the dynamic stability of the steady-state solution to small perturbations, or the solution of time-dependent problems that are near a steady state. Such *quasisteady* problems require a time-accurate approach that can also handle steady states well. One approach is presented in Section 17.14.

Fractional-step method can often be used to successfully compute steady-state or quasisteady solutions, but several issues arise. As mentioned above, the steady state results from a balance (cancellation) between two dynamic processes that are a handled separately in a fractional-step method. In some cases the method may not even converge, but instead will oscillate in time near the correct solution. This can happen if a high-resolution method with limiter functions is used for the hyperbolic part, since the limiter depends on the solution and effectively switches between different methods based on the behavior of the solution.

Even when the method converges, the numerical steady state obtained will typically depend on the time step used. This is rather unsatisfying, since the steady solution depends only on x and so we would like the numerical solution generated by a particular method to depend only on Δx. By contrast, unsplit methods can often be developed in which the steady state is independent of Δt. See Exercise 17.4 for one example.

17.9 Boundary Conditions for Fractional-Step Methods

When a fractional-step method is used, we typically need to impose boundary conditions in the hyperbolic step of the procedure. We may also need to impose boundary conditions in the source term step(s) if the source terms involve spatial derivatives of q – for example, if these are diffusion terms, then we are solving the diffusion equation, which requires boundary conditions at each boundary. The boundary conditions for the original PDE must be used to determine any boundary conditions needed for the fractional steps, but the connection between these is often nontrivial.

As a simple example, consider the advection–reaction equation (17.7) with the constant boundary data $q(0,t) = g_0(t) \equiv 1$, which results in the steady-state solution (17.44) for large t (with $C = 1$). Suppose we use a fractional-step method with the Godunov splitting, in which we first solve the advection equation $q_t + \bar{u} q_x = 0$ over time Δt and then the ODE $q_t = -\beta q$ over time Δt. Moreover, suppose we choose Δt so that $\bar{u}\,\Delta t/\Delta x = 1$ and the advection equation is solved exactly via

$$Q_i^* = Q_{i-1}^n, \tag{17.46}$$

and then we also solve the ODE exactly via

$$Q_i^{n+1} = e^{-\beta\,\Delta t} Q_i^*. \tag{17.47}$$

These steps can be combined to yield

$$Q_i^{n+1} = e^{-\beta\,\Delta t} Q_{i-1}^n. \tag{17.48}$$

For this simple problem we observed in Section 17.2.2 that there is no splitting error, and hence this procedure should yield the exact solution. If Q_{i-1}^n is the exact cell average of $q(x, t_n)$ over cell $\mathcal{C}_{i-1}$, then Q_i^{n+1} will be the exact cell average of $q(x, t_{n+1})$ over cell $\mathcal{C}_i$.

But to determine Q_1^{n+1} we must also use the boundary conditions. To implement the step (17.46) at $i = 1$ we must first specify a ghost cell value Q_0^n as described in Chapter 7. (Note that the step (17.47) does not require any boundary data, since we are solving an ODE within each grid cell.)

As a first guess at the value Q_0^n, we might follow the discussion of Section 7.2.2 and use the integral of (7.8). This would give $Q_0^n = 1$, since the specified boundary condition is independent of time. It appears that we have specified the exact ghost-cell value, and we know that the method being used in the interior is exact, and yet this combination will *not* produce the exact solution numerically. The value Q_1^{n+1} will not be the exact cell average of $q(t, t_{n+1})$ over the cell $\mathcal{C}_1$. The computed value will be

$$Q_1^{n+1} = e^{-\beta \Delta t} = e^{-(\beta/\bar{u})\Delta x} = 1 - (\beta/\bar{u})\Delta x + \mathcal{O}(\Delta x^2), \tag{17.49}$$

while the true cell average of the solution (17.44) is easily computed to be

$$\frac{1}{\Delta x}\int_{\mathcal{C}_1} e^{-(\beta/\bar{u})x}\,dx = -\frac{\bar{u}}{\beta\,\Delta x}\left(e^{-(\beta/\bar{u})\Delta x} - 1\right) = 1 - \frac{1}{2}(\beta/\bar{u})\Delta x + \mathcal{O}(\Delta x^2). \tag{17.50}$$

The numerical value (17.49) is too small by $\mathcal{O}(\Delta x)$. In later time steps this error will propagate downstream, and eventually the entire solution will have an $\mathcal{O}(\Delta x)$ error.

The reason for this error is made clear in Figure 17.5. Figure 17.5(a) shows the steady-state solution, which is used as initial data, along with the function obtained after time Δt if we solve the advection equation alone with the boundary condition $g_0(t) = 1$. Figure 17.5(b) shows the results if we now solve the decay equation using the advected solution as data. Away from the boundary the decay exactly cancels the apparent growth due to the advection, and the exact steady-state solution is recovered. Near the boundary the use of the constant inflow boundary condition $g_0(t) = 1$ leads to the wrong profile.

It is clear how to fix this once we realize that the boundary condition $q(0, t) = 1$ is the proper boundary condition for the full equation $q_t + \bar{u}q_x = -\beta q$, but is *not* the proper boundary condition for the pure advection equation $q_t^* + \bar{u}q_x^* = 0$ that is being solved by

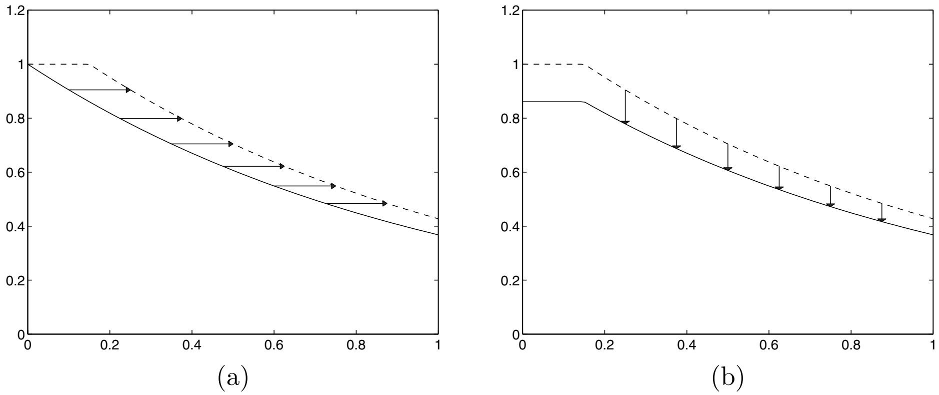


Fig. 17.5. The effect of incorrect boundary conditions in a fractional-step method. (a) The steady-state solution $q(x) = e^{-(\beta/\bar{u})x}$ is shown as the solid line. After advection by a distance $\bar{u}\,\Delta t$ to the right with $q^*(0, t) = 1$, the solution is shown as the dashed line. (b) The advected solution decays by $e^{-\beta\Delta t}$ and an error is apparent near the boundary.

the upwind method in the first step of the fractional-step method. If we denote the solution to this equation by $q^*(x, t)$ for $t \geq t_n$, then this function is different from $q(x, t)$ for $t > t_n$ and requires different boundary conditions. Figure 17.4(b) suggests what the correct boundary conditions are. We would like $q^*(x, t_n + \Delta t)$ to be the function $e^{-(\beta/\bar{u})(x-\bar{u}\,\Delta t)} = e^{\beta\,\Delta t}e^{-(\beta/\bar{u})x}$ for all $x \geq 0$, so that after the decay step we will recover $e^{-(\beta/\bar{u})x}$ for all $x \geq 0$. To obtain this we clearly need to impose a boundary condition on $q^*(0, t)$ that is growing with t,

$$q^*(0, t) = e^{\beta(t-t_n)} \equiv g_0^*(t),$$

and it is this function that should be used in determining the ghost-cell value Q_0^n instead of the original boundary condition $g_0(t) = 1$. Note that if we evaluate the integral from (7.8) using this function $g_0^*(t)$, we obtain

$$Q_0^n = \frac{\bar{u}}{\beta\,\Delta x}\left(e^{(\beta/\bar{u})\Delta x} - 1\right). \tag{17.51}$$

Using this boundary value in the formulas (17.47) and (17.48) results in

$$Q_1^{n+1} = e^{-\beta\,\Delta t}Q_0^n = \frac{\bar{u}}{\beta\,\Delta x}(1 - e^{-\beta\Delta x}), \tag{17.52}$$

which is exactly the value (17.50).

For the advection–decay equation with a more general time-dependent boundary condition $q(0, t) = g_0(t)$, the proper boundary condition to use in the advection step is

$$q^*(0, t) = g_0^*(t) = e^{\beta(t-t_n)}g_0(t).$$

The integral (7.8) can perhaps be evaluated exactly using this function in place of g_0, or an approximation such as (7.9) could again be used, which would result in

$$Q_0^n = e^{\beta\,\Delta t/2}g_0\left(t_n + \frac{\Delta x}{2\bar{u}}\right). \tag{17.53}$$

The proper value for the ghost-cell value Q_{-1}^n can be found similarly if a second ghost cell is needed (for a method with a wider stencil).

For the simple example considered above it was easy to determine the correct boundary conditions for q^* based on our knowledge of the exact solution operators for the advection and decay problems. For other problems it may not be so easy to determine the correct boundary conditions, but an awareness of the issues raised here can often help in deriving better boundary conditions than would be obtained by simply using the boundary conditions from the original equation. Since the required modification to the boundary conditions is typically $\mathcal{O}(\Delta t)$, as in the above example, this can make a significant difference in connection with methods that are second-order accurate or better. A more general procedure for deriving the proper intermediate boundary conditions for a linear hyperbolic equation is discussed in [276].

17.10 Stiff and Singular Source Terms

Most of the source terms that have appeared so far have been bounded functions corresponding to a source that is distributed in space. In some problems source terms naturally arise that are concentrated at a single point, a delta function, or at least are concentrated over a very small region compared to the size of the domain. In some cases this may be an external source depending explicitly on x that is spatially concentrated. An example of this nature was presented in Section 16.3.1, where we considered an advection equation with a delta function source of tracer. We consider another problem of this type in Section 17.11.

In other cases the source $\psi(q)$ depends only on the solution and yet the solution naturally develops structures in which the source terms are nonzero (and very large) only over very small regions in space. For example, this often happens if the source terms model chemical reactions between different species (reacting flow) in cases where the reactions happen on time scales much faster than the fluid dynamic time scales. Then solutions can develop thin *reaction zones* where the chemical-kinetics activity is concentrated. Such problems are said to have *stiff source terms*, in analogy with the classical case of stiff ODEs. Stiffness is common in kinetics problems. Reaction rates often vary by many orders of magnitude, so that some reactions occur on time scales that are very short compared to the time period that must be studied. One classic example of a stiff reacting flow problem is a *detonation wave*; an explosion in which a flammable gas burns over a very thin reaction zone that moves through the unburned gas like a shock wave, but with a more complicated structure. (See, for example, [92], [136], [156].) The thin reaction zone can be idealized as a delta-function source term that moves with the detonation wave. Some simpler examples are studied in this chapter. We will see that singular source terms lead to a modification of the Rankine–Hugoniot jump conditions that determine the structure and speed of propagating discontinuities.

17.11 Linear Traffic Flow with On-Ramps or Exits

As an illustration of a hyperbolic equation with a singular source term, consider traffic flow on a one-lane highway with on-ramps and exits where cars can enter or leave the highway. Then the total number of cars on the highway is not conserved, and instead there are sources and sinks. The corresponding source terms in the equation are delta functions with positive strength at the locations of the on-ramps and negative strength at exits.

As an example, consider a single on-ramp with a flux D at some point x_0, so that the source term is

$$\psi(x) = D\delta(x - x_0). \tag{17.54}$$

To begin with, suppose the traffic is sufficiently light that it moves at some constant speed $\bar{u}$ independent of the density q. Then the traffic flow is modeled by an advection equation as in Section 9.4.2 with the addition of the source term (17.54),

$$q_t + \bar{u}q_x = D\delta(x - x_0). \tag{17.55}$$

This is exactly the problem considered in Section 16.3.1, and the Riemann solution has a jump from q_l to $q_m = q_l + D/\bar{u}$ at $x = x_0$ (resulting from the source) and then a jump from q_m to q_r at $x = \bar{u}t$ (resulting from the initial data, and moving downstream with the traffic).

17.12 Rankine–Hugoniot Jump Conditions at a Singular Source

The jump in q at the on-ramp can be derived from a more general formula, an extension of the Rankine–Hugoniot jump condition to the case where there is a singular source moving with the jump. This formula will be needed to study the nonlinear traffic flow problem.

Consider a general conservation law coupled with a delta-function source term moving at some speed $s(t)$,

$$q_t + f(q)_x = D\,\delta(x - X(t)), \tag{17.56}$$

where $X'(t) = s(t)$. We will compute jump conditions at $X(t)$ using the same procedure as in Section 11.8 for the homogeneous equation. The differential equation (17.56) results from an integral conservation law that, over a small rectangular region such as the one shown in Figure 11.7, has the form (for $s < 0$ as in the figure)

$$\begin{aligned}\int_{x_1}^{x_1+\Delta x} q(x, t_1 + \Delta t)\,dx - \int_{x_1}^{x_1+\Delta x} q(x, t_1)\,dx \\ = \int_{t_1}^{t_1+\Delta t} f(q(x_1, t))\,dt - \int_{t_1}^{t_1+\Delta t} f(q(x_1 + \Delta x, t))\,dt \\ + \int_{t_1}^{t_1+\Delta t} \int_{x_1}^{x_1+\Delta x} D\,\delta(x - X(t))\,dx\,dt.\end{aligned} \tag{17.57}$$

Note that over this time interval the point $X(t)$ always lies between x_1 and $x_1 + \Delta x$, so that

$$\int_{x_1}^{x_1+\Delta x} D\,\delta(x - X(t))\,dx = D,$$

and hence the equation (17.57) can be approximated by

$$\Delta x\, q_r - \Delta x\, q_l = \Delta t\, f(q_l) - \Delta t\, f(q_r) + \Delta t D + \mathcal{O}(\Delta t^2). \tag{17.58}$$

Using $\Delta x = -s\,\Delta t$, dividing by $-\Delta t$, and taking the limit as $\Delta t \to 0$ gives

$$s(q_r - q_l) = f(q_r) - f(q_l) - D. \tag{17.59}$$

This is identical to the Rankine–Hugoniot jump condition (11.20) but with an additional term resulting from the singular source.

Note that if the source term $\psi(x, t)$ were a bounded function rather than a delta function, then the source term in (17.57) would be

$$\begin{aligned}\int_{t_1}^{t_1+\Delta t} \int_{x_1}^{x_1+\Delta x} \psi(x, t)\,dx\,dt &\approx \Delta t \Delta x\, \psi(x, t) \\ &\approx \Delta t^2 s\, \psi(x_1, t_1).\end{aligned} \tag{17.60}$$

After dividing by $-\Delta t$ this would still be $\mathcal{O}(\Delta t)$ and would vanish as $\Delta t \to 0$. Hence a bounded source term does not change the Rankine–Hugoniot jump condition (11.20) at a discontinuity, since its contribution at any single point is negligible. A delta-function source term makes a nontrivial contribution at a single point and hence appears in the jump condition at that point.

Example 17.6. Consider again the on-ramp problem of Section 17.11, modeled by the equation (17.55). In this case $s = 0$ and $f(q) = \bar{u}q$, so that at $x = x_0$, where q jumps from q_l to q_m, the jump condition (17.59) yields

$$q_m - q_l = D/\bar{u}.$$

Note that this makes sense physically: D measures the flux per unit time of cars onto the highway, but the larger $\bar{u}$ is, the more widely spaced these cars are in the existing traffic, and hence the smaller the effect on the density.

Note that this Riemann solution has a similar structure to the Riemann solution illustrated in Figure 9.3 for the the variable-coefficient advection equation

$$q_t + (u(x)q)_x = 0 \tag{17.61}$$

in the case where $u(x)$ is discontinuous with a single jump at x_0. In fact there is a connection between the two, since (17.61) can be rewritten as

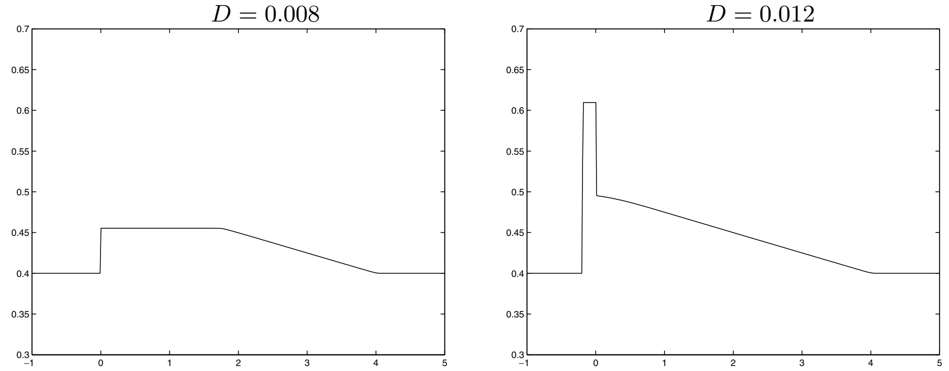
$$q_t + u(x)q_x = -u'(x)q.$$

This is the color equation with a source term. For the case described in Figure 9.3, $u(x)$ is piecewise constant and hence $u'(x)$ becomes a delta function at x_0.

17.13 Nonlinear Traffic Flow with On-Ramps or Exits

For larger traffic densities a nonlinear traffic model must be used, and again a source term representing incoming cars at an on-ramp can be included. Figure 17.6 shows examples using the traffic flow model of Section 11.1 with velocity function (11.5). The initial density was $q = 0.4$ everywhere and a source of strength D is introduced at $x_0 = 0$ starting at time $t = 0$. In Figure 17.6(a) the source strength is small enough that the structure is essentially the same as it would be in a linear problem: congestion is seen only downstream from the on-ramp, and cars speed up again through a rarefaction wave. Figure 17.6(b) shows the same problem with a slightly larger value of D, in which case the structure is quite different. Since velocity varies with density in the nonlinear model, when the flux of cars onto the highway is too large a traffic-jam shock wave forms and moves upstream from the on-ramp. Note that even if the flux at the on-ramp is now reduced or eliminated this traffic jam will continue to propagate upstream and disrupt traffic. This is one reason that some on-ramps are equipped with traffic lights allowing only "one car per green" to insure that the flux never rises above a certain critical value.

For this model the structure of the Riemann solution can be explicitly determined (see Exercise 17.6). If $q_l > 0.5$, then characteristic signals travel upstream and a traffic-jam shock



(a) (b)

Fig. 17.6. Nonlinear traffic flow with a source term at an on-ramp. (a) The source strength $D = 0.008$ is sufficiently low that upstream traffic is unaffected. (b) For greater source strength $D = 0.012$ a traffic jam moves upstream from the on-ramp. Shown at $t = 20$. `[claw/book/chap17/onramp]`

will form for any $D > 0$. If $q_l < 0.5$, then a shock forms whenever $D > (1 - 2q_l)^2/4$. For the example in Figure 17.6, this cutoff is at $D = 0.01$.

17.14 Accurate Solution of Quasisteady Problems

In Section 17.8 we observed that fractional-step methods may not be well suited to problems near a steady state, where $f(q)_x$ and $\psi(q)$ are each large in magnitude but nearly cancel out. An alternative unsplit method can be formulated by discretizing the source term as a sum of delta function sources located at the cell interfaces,

$$q_t + f(q)_x = \sum_i \Delta x \, \Psi_{i-1/2}(t)\delta\big(x - x_{i-1/2}\big). \tag{17.62}$$

At time t_n we then have a Riemann problem at $x_{i-1/2}$ for the equation

$$q_t + f(q)_x = \Delta x \, \Psi^n_{i-1/2}\delta\big(x - x_{i-1/2}\big), \tag{17.63}$$

with data Q_{i-1} and Q_i. Since the delta-function source is at a fixed location, the solution will consist of propagating waves along with a jump in q at $x_{i-1/2}$, a similar structure to what is observed when for a conservation law with a spatially varying flux function, as discussed in Section 16.4.

Now suppose we use a linearized approximate Riemann solver of the form discussed in Section 15.3 to replace the flux $f(q)$ by $\hat{A}_{i-1/2}q$, where $\hat{A}_{i-1/2}$ is an approximate Jacobian matrix determined by the data Q_{i-1} and Q_i. Then we have a Riemann problem for the equation

$$q_t + \hat{A}_{i-1/2}q_x = \Delta x \, \Psi_{i-1/2}\delta\big(x - x_{i-1/2}\big). \tag{17.64}$$

The solution will consist of waves propagating with speeds $s^p_{i-1/2} = \hat{\lambda}^p_{i-1/2}$, the eigenvalues of $\hat{A}_{i-1/2}$, and an additional discontinuity in q at $x_{i-1/2}$ (propagating at speed 0) arising from the delta-function source. Instead of a single state $Q^{\Downarrow}_{i-1/2}$ at $x_{i-1/2}$ in the Riemann solution, there will be two states $Q_l^{\Downarrow}$ and $Q_r^{\Downarrow}$ just to the left and right of this ray, satisfying the Rankine–Hugoniot relation (17.59) with $s = 0$,

$$\hat{A}_{i-1/2} Q_r^{\Downarrow} - \hat{A}_{i-1/2} Q_l^{\Downarrow} = \Delta x\, \Psi_{i-1/2}. \tag{17.65}$$

These states must be related to Q_{i-1} and Q_i by

$$\hat{A}_{i-1/2} Q_l^{\Downarrow} - \hat{A}_{i-1/2} Q_{i-1} = \sum_{p:s_{i-1/2}^p < 0} s_{i-1/2}^p \mathcal{W}_{i-1/2}^p \tag{17.66}$$

and

$$\hat{A}_{i-1/2} Q_i - \hat{A}_{i-1/2} Q_r^{\Downarrow} = \sum_{p:s_{i-1/2}^p > 0} s_{i-1/2}^p \mathcal{W}_{i-1/2}^p, \tag{17.67}$$

where $\mathcal{W}_{i-1/2}^p = \alpha_{i-1/2}^p \hat{r}_{i-1/2}^p$ is the pth wave, which is proportional to the eigenvector $\hat{r}_{i-1/2}^p$ of $\hat{A}_{i-1/2}^p$. Adding (17.66) and (17.67) together and using (17.65) allows us to eliminate $Q_{l,r}^{\Downarrow}$ and obtain

$$\hat{A}_{i-1/2}(Q_i - Q_{i-1}) - \Delta x\, \Psi_{i-1/2} = \sum_{p=1}^{m} s_{i-1/2}^p \mathcal{W}_{i-1/2}^p. \tag{17.68}$$

If the Roe solver of Section 15.3.2 is used, then the matrix $\hat{A}_{i-1/2}$ satisfies (15.18), and so (17.68) becomes

$$f(Q_i) - f(Q_{i-1}) - \Delta x\, \Psi_{i-1/2} = \sum_{p=1}^{m} s_{i-1/2}^p \mathcal{W}_{i-1/2}^p. \tag{17.69}$$

In order to determine the waves $\mathcal{W}_{i-1/2}^p$ in the Riemann solution, we need only decompose $f(Q_i) - f(Q_{i-1}) - \Delta x \Psi_{i-1/2}$ into eigenvectors as

$$f(Q_i) - f(Q_{i-1}) - \Delta x\, \Psi_{i-1/2} = \sum_{p=1}^{m} \beta_{i-1/2}^p \hat{r}_{i-1/2}^p, \tag{17.70}$$

and then set $\mathcal{W}_{i-1/2}^p = \alpha_{i-1/2}^p \hat{r}_{i-1/2}^p$, where

$$\alpha_{i-1/2}^p = \beta_{i-1/2}^p / s_{i-1/2}^p \quad \text{for } s_{i-1/2}^p \neq 0. \tag{17.71}$$

Godunov's method and related high-resolution methods, when implemented in the wave-propagation form, only require the waves propagating at nonzero speeds, and these can all be obtained by this procedure. Alternatively, the formulation of Section 15.5 can be used to implement the method directly in terms of the waves $\mathcal{Z}_{i-1/2}^p = \beta_{i-1/2}^p \hat{r}_{i-1/2}^p$, with the simple modification that these waves are now defined by solving (17.70) instead of (15.67). As in the discussion of Section 15.5, this procedure can also be used for choices of $\hat{r}^p$ other than the eigenvectors of the Roe matrix.

This method is generally easy to implement by a simple modification of the Riemann solver for the homogeneous equation. It is not formally second-order accurate in general, and may not perform as well as fractional-step methods for some time-dependent problems that are far from steady state.

The method can be greatly advantageous, however, for quasisteady problems, where $f(q)_x \approx \psi(q)$. In this case we expect

$$\frac{f(Q_i) - f(Q_{i-1})}{\Delta x} \approx \Psi_{i-1/2}, \tag{17.72}$$

and hence the left-hand side of (17.69) will be near 0. The waves resulting from the eigendecomposition will thus have strength near zero, and will cause little change in the solution. These waves will model the deviation from steady state, and it is precisely this information that should be propagated, and to which wave limiters should be applied. Moreover, a numerical steady-state solution computed with this method will satisfy (17.72) with equality. If the source term is appropriately discretized then smooth steady-state solutions will be computed with second-order accuracy even though the transient behavior may not be formally second-order accurate.

A different wave-propagation method with similar features was proposed in [284], but the approach just presented seems to be more robust as well as much easier to implement; see [18] for more discussion. Some other related methods can be found in the references of Section 17.2.1.

17.15 Burgers Equation with a Stiff Source Term

In the remainder of this chapter we consider problems having source terms that do not appear to contain delta functions, but that are typically close to zero over most of the domain while being very large over thin *reaction zones* that dynamically evolve as part of the solution. Such source terms can often be approximated by delta functions, but their location and strength is generally not known *a priori*.

As a simple but illustrative example, consider the Burgers equation with a source term,

$$u_t + \frac{1}{2}(u^2)_x = \psi(u), \tag{17.73}$$

where

$$\psi(u) = \frac{1}{\tau}u(1-u)(u-\beta), \tag{17.74}$$

with $\tau > 0$ and $0 < \beta < 1$. If we consider the ODE

$$u'(t) = \psi(u(t)) \tag{17.75}$$

alone, we see that $u = 0,\ \beta, 1$ are equilibrium points. The middle point $u = \beta$ is an unstable equilibrium, and from any initial data $\mathring{u} \neq \beta$, u asymptotically approaches one of the other equilibria: $u \to 0$ if $\mathring{u} < \beta$ or $u \to 1$ if $\mathring{u} > \beta$. The parameter τ determines the time scale over which u approaches an equilibrium.

For τ very small, any initial data $\mathring{u}(x)$ supplied to the equation (17.73) will rapidly approach a step function with values 0 and 1 (and near-discontinuous behavior where $\mathring{u}(x)$ passes through β) on a much faster time scale than the hyperbolic wave propagation. To see how the solution then evolves, it suffices to consider the case of a Riemann problem with jump from value 0 to 1 or from 1 to 0.

If $u_l = 1$ and $u_r = 0$, then the Burgers equation with no source gives a shock wave moving with the Rankine–Hugoniot speed $1/2$, by (11.23). The source term is then identically zero and has no effect. More general initial data $\mathring{u}(x)$ satisfying $\mathring{u}(x) > \beta$ for $x < 0$ and $\mathring{u}(x) < \beta$ for $x > 0$ would rapidly evolve to this situation and give a shock traveling with speed $1/2$ for any value of $\beta \in (0, 1)$.

The situation is much more interesting in the case where $u_l = 0$ and $u_r = 1$, as studied in [280]. In this case Burgers' equation gives a rarefaction wave that spreads the initial discontinuity out through all intermediate values. The source term opposes this smearing and sharpens all values back towards 0 or 1. These competing effects balance out and result in a smooth solution that rapidly approaches a traveling wave that neither smears nor sharpens further, but simply propagates with some speed s:

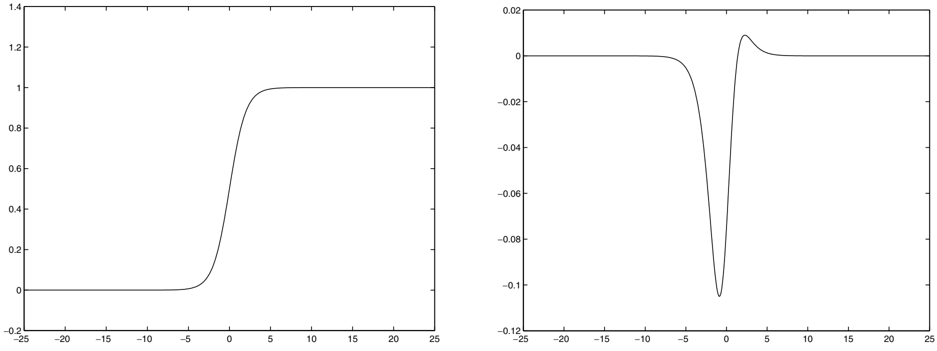
$$u(x, t) = w\left(\frac{x - st}{\tau}\right). \tag{17.76}$$

We will see below that $s = \beta$.

The shape of this profile is shown in Figure 17.7(a). The width of the transition zone in $u(x, t)$ is $\mathcal{O}(\tau)$, and for small τ this appears similar to a viscous shock profile with small viscosity, though the competing effects that produce the steady profile are different in that case. As we will see in the next section, computing such a solution on a grid where the structure is not well resolved (e.g., for $\Delta x > \tau$) can be more difficult than correctly approximating a shock wave.

The shape of the profile w and the speed s can be determined by inserting the form (17.76) into (17.73), yielding

$$-sw' + ww' = w(1 - w)(w - \beta), \tag{17.77}$$



(a) (b)

Fig. 17.7. (a) Traveling-wave solution $w(\xi)$ to (17.73). (b) The source term $\psi(w(\xi))$ of (17.74). Shown for $\beta = 0.6$.

which gives the ODE

$$w' = \frac{w(1-w)(w-\beta)}{w-s}. \tag{17.78}$$

We also require $w(-\infty) = 0$ and $w(+\infty) = 1$. The equation (17.78) has a solution with these limiting values only if $s = \beta$, since otherwise w cannot cross the unstable equilibrium value β. When $s = \beta$ we can cancel this term in (17.78) and obtain the logistic equation

$$w' = w(1-w), \tag{17.79}$$

with solutions of the form

$$w(\xi) = \frac{e^{\xi}}{1+e^{\xi}} = \frac{1}{2}[1 + \tanh(\xi/2)], \tag{17.80}$$

as shown in Figure 17.7(a).

The propagation speed is $s = \beta$. Unlike the case $u_l = 1,\ u_r = 0$, this speed depends on the structure of the source term. This leads to numerical difficulties if we try to solve the problem on an underresolved grid, as discussed in the next section. More generally, if we replace $f(u) = u^2/2$ by a more general flux function, the propagation speed of the resulting traveling wave will be $s = f'(\beta)$, as shown in [129].

We can relate the result just found to the discussion of singular source terms earlier in this chapter by observing that the source term $\psi(w)$ will be essentially zero except in the transition region, where its magnitude is $\mathcal{O}(1/\tau)$, as seen in Figure 17.7(b). Since this region has width $\mathcal{O}(\tau)$, this suggests that the source terms approximate a delta function as $\tau \to 0$.

The magnitude of the limiting delta function can be found by integrating $\psi(w((x-st)/\tau))$ over all x and taking the limit as $\tau \to 0$, although in fact this value is independent of τ for the wave form w. We can use (17.79) to rewrite $\psi(w(\xi))$ as

$$\begin{aligned} \psi(w(\xi)) &= \frac{1}{\tau} w'(\xi)[w(\xi) - \beta] \\ &= \frac{1}{\tau}\frac{d}{d\xi}\left(\frac{1}{2}[w(\xi) - \beta]^2\right), \end{aligned} \tag{17.81}$$

and hence

$$\begin{aligned} \int_{-\infty}^{\infty} \psi(w(x-st)/\tau)\,dx &= \tau \int_{-\infty}^{\infty} \psi(w(\xi))\,d\xi \\ &= \frac{1}{2}[w(\xi) - \beta]^2 \Big|_{-\infty}^{+\infty} \\ &= \frac{1}{2} - \beta. \end{aligned} \tag{17.82}$$

This value is independent of τ and gives the strength D of the delta-function source observed in the limit $\tau \to 0$,

$$D = \frac{1}{2} - \beta. \tag{17.83}$$

Using the modified Rankine–Hugoniot relation (17.59) results in the speed

$$s = \frac{1}{2} - D = \beta$$

for the limiting jump discontinuity, which is consistent with the speed of the traveling wave, as we should expect. We see that the source term has a nontrivial effect even in the limit $\tau \to 0$ whenever $\beta \neq 1/2$.

Note from Figure 17.7(b) that the source term is negative where $w < \beta$ and positive where $w > \beta$. When $\beta = 1/2$ these two portions exactly cancel and there is no net source term, so the propagation speed agrees with what is expected from the conservation law alone. When $\beta \neq 1/2$ there is a net source or sink of u in the transition zone that affects the speed at which the front propagates. This same effect is seen in many reacting flow problems.

17.16 Numerical Difficulties with Stiff Source Terms

A hyperbolic equation with a source term is said to be *stiff* if the wave-propagation behavior of interest occurs on a much slower time scale than the fastest times scales of the ODE $q_t = \psi(q)$ arising from the source term. An example is a detonation wave arising when gas dynamics is coupled with the chemical kinetics of combustion. Detonation waves can travel rapidly through a combustible gas, but even these fast waves are many orders of magnitude slower than the time scales of some of the chemical reactions appearing in the kinetic source terms. It would often be impossible to simulate the propagation of such waves over distances of physical interest if it were necessary to fully resolve the fastest reactions.

As a simpler example, consider the Burgers equation (17.73) with source (17.74). As we have seen in the previous section, when τ is very small a traveling-wave solution looks essentially like a discontinuity from $u_l = 0$ to $u_r = 1$ propagating at speed $\beta = \mathcal{O}(1)$. Suppose we want to approximate this with $\tau = 10^{-10}$, say, over a domain and time interval that are $\mathcal{O}(1)$. Then the transition from u_l to u_r takes place over a zone of width on the order of 10^{-10}. We will certainly want to use $\Delta x \gg \tau$, and we can't hope to resolve the structure of the traveling wave, only the macroscopic behavior (as in shock capturing). We also do not want to take $\Delta t = \mathcal{O}(\tau)$, but rather want to take $\Delta t = \mathcal{O}(\Delta x)$ based on the CFL restriction for the hyperbolic problem. We hope that if the wave of interest moves less than one grid cell each time step, we will be able to capture it accurately, with little numerical smearing and the physically correct velocity.

The classical theory of numerical methods for stiff *ordinary* differential equations can be found in many sources, e.g., [145], [178], [253]. Recall that an ODE is called stiff if we are trying to compute a particular solution that is varying much more slowly than the fastest time scales inherent in the problem. Typically this means that the solution of interest is evolving on some *slow manifold* in state space and that perturbing the solution slightly off this manifold will result in a rapid transient response, bringing the solution back the manifold, followed by slow evolution once again. Stiff problems often arise in chemical kinetics, where the rates of different reactions can vary by many orders of magnitude. If the fastest reactions are essentially in equilibrium, then the concentrations will vary slowly on time scales governed by slower reactions, and the solution is evolving on a slow manifold.

If the system is perturbed, say by injecting additional reactants, then fast transient behavior may result over a short period of time as the faster reactions reach a new equilibrium.

A simpler example of a stiff ODE is the equation (17.75) with $\tau \ll T$ in (17.74), where T is some reference time over which we are interested in the solution. For most initial data there will be fast transient decay to one of the slow manifolds $u \equiv 0$ or $u \equiv 1$. Perturbing away from one of these values results in a fast transient. An even simpler example is the equation

$$u'(t) = -u(t)/\tau, \tag{17.84}$$

where $u \equiv 0$ is the slow manifold.

Numerically, stiff ODEs are problematical because we typically want to take "large" steps whenever the solution is "slowly" evolving. However, the numerical method is not exact and hence is constantly perturbing the solution. If these perturbations take the solution off the slow manifold, then in the next time step the solution will have a rapid transient and the desired time step may be much too large. In particular, if an explicit method is used, then the stability restriction of the method will generally require choosing the time step based on the fastest time scale present in the problem, even if we are hoping we do not need to resolve this scale. Luckily many *implicit* methods have much better stability properties and allow one to choose the time step based on the behavior of the solution of interest.

As we will see, solving stiff hyperbolic equations can be even more challenging than solving stiff ODEs. This difficulty arises largely from the fact that in a stiff hyperbolic equation the fastest reactions are often *not* in equilibrium everywhere. In thin regions such as the transition zone of the problem considered in Section 17.15 (see Figure 17.7(b)), or the reaction zone of a detonation wave, there is a fast transient behavior constantly taking place. For some problems it really appears necessary to resolve this zone in order to obtain good solutions, in which case adaptive mesh refinement is often required to efficiently obtain good results. In other cases we can achieve our goal of solving stiff problems on underresolved grids. A number of techniques have been developed for various problems, and here we will primarily illustrate some of the difficulties.

For stiff source terms we will again concentrate on the use of fractional-step methods (though more sophisticated approaches may not use splitting). The *Godunov splitting* applied to $q_t + f(q)_x = \psi(q)$ would simply alternate between solving the following two problems over time step Δt:

$$\text{Problem A:} \quad q_t + f(q)_x = 0, \tag{17.85}$$

$$\text{Problem B:} \quad q_t = \psi(q). \tag{17.86}$$

The first thing to observe is that the ODE of (17.86) is going to be stiff, and so we must use an appropriate method for this step in going from Q_i^* to Q_i^{n+1} in the ith cell. One popular class of stiff solvers are the BDF methods (backward differentiation formulas). These are linear multistep methods and require at least two previous time levels to get second-order accuracy or better. As discussed in Section 17.6, this is a problem in the context of a fractional-step method, since we only have one value Q_i^* that we can use.

The trapezoidal method (17.41) can often be effectively used for stiff ODEs, but may fail miserably for hyperbolic equations with stiff source terms. Figure 17.8 shows a sample

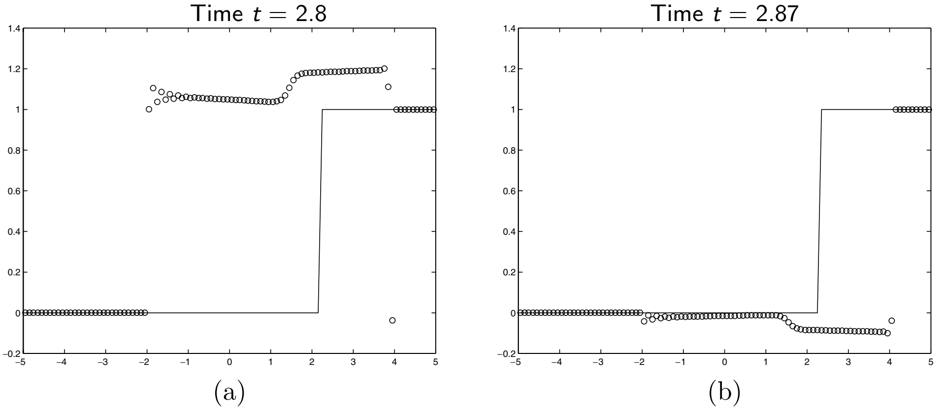


Fig. 17.8. Numerical solution to the Burgers equation with a stiff source term using the trapezoidal method for the source term: (a) after 40 time steps, (b) after 41 time steps. [claw/book/chap17/stiffburgers]

computation on the Burgers-equation example of Section 17.15, with $\beta = 0.8$, $\tau = 10^{-5}$, $\Delta x = 0.1$, and $\Delta t = 0.07$. The initial data was $u = 0$ for $x < 0$ and $u = 1$ for $x > 0$. Figure 17.8(a) shows the solution after 40 time steps, while Figure 17.8(b) shows the solution one time step later. The solution oscillates in time between these two different basic shapes with a set of waves propagating at unphysical speeds.

This behavior arises from the fact that the trapezoidal method is not *L-stable* (see [253]). If we start on the slow manifold, this method does a good job of keeping the numerical solution on the slow manifold even with time steps that are large relative to the faster scales. But for initial data that has an initial fast transient, the method yields oscillations. This is easy to see for the simple ODE (17.84). In this case the "slow manifold" is simply $u \equiv 0$, and starting with any other data gives exponentially fast decay of the true solution towards this state, with rate $1/\tau$. On this problem the trapezoidal method yields

$$U^{n+1} = \left(\frac{1 - \frac{1}{2}\Delta t/\tau}{1 + \frac{1}{2}\Delta t/\tau}\right) U^*. \tag{17.87}$$

If $U^* = 0$ then $U^{n+1} = 0$ also and we stay on the slow manifold. But if $U^* \neq 0$ and $\Delta t/\tau \gg 1$ then $U^{n+1} \approx -U^*$. The "amplification factor" in (17.87) approaches -1 as $-\Delta t/\tau \to -\infty$. Rather than the proper decay, we obtain an oscillation in time unless the transient is well resolved.

For the Burgers equation (17.73) with a stiff source term, nonequilibrium data is constantly being introduced by the averaging process in solving the conservation law near any front. This sets up an oscillation in time, as observed in Figure 17.8.

The trapezoidal method is an A-stable method. In fact, its stability region is exactly the left half plane. The fact that the boundary of the stability region is the imaginary axis, and hence passes through the point at infinity on the Riemann sphere, is responsible for the observed undesirable behavior, since in solving the stiff problem we are interested in letting $-\Delta t/\tau \to -\infty$.

An *L-stable* method is one for which the point at infinity is in the interior of the stability region, and hence the amplification factor approaches something less than 1 in magnitude as $-\Delta t/\tau \to -\infty$. (L-stability is usually defined to also require A-stability, but we use this looser definition for convenience.) The BDF methods have this property. The one-step BDF method is the backward Euler method, which for the equation (17.84) is simply

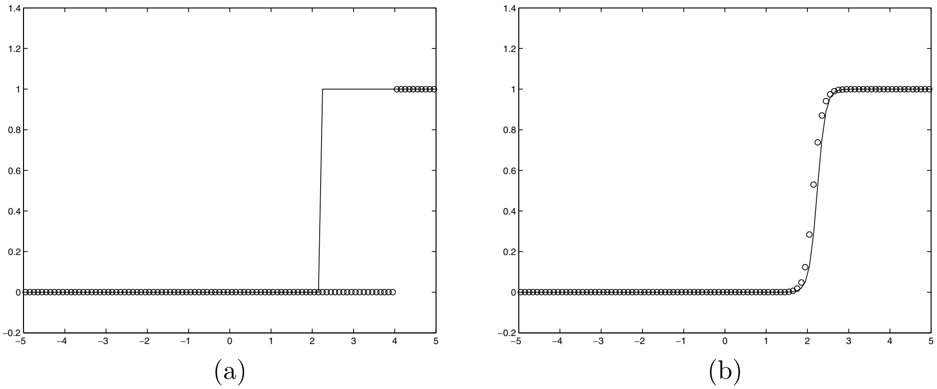
$$U^{n+1} = U^* - \frac{\Delta t}{\tau} U^{n+1} \quad \Longrightarrow \quad U^{n+1} = \left(\frac{1}{1 + \Delta t/\tau} \right) U^*. \tag{17.88}$$

Note that $(1 + \Delta t/\tau)^{-1} \to 0$ as $\Delta t/\tau \to -\infty$, and so this method can be used on stiff problems even when the initial data is not on the slow manifold. The backward Euler method is only first-order accurate. In the present context, this does not really matter, since we expect the source terms to be active only over thin regions where there are fast transients that we cannot resolve with any accuracy anyway. What we primarily require is the L-stability property.

In some situations we may want to use a second-order method that is L-stable so that we can obtain good accuracy of source terms in regions where they are smooth (or where transients are well resolved) and also avoid oscillations in regions of stiffness. In the context of a fractional-step method we must use a one-step method, and one possible choice is the TR-BDF2 method (17.43) described in Section 17.6, which is L-stable. Figure 17.9 shows results obtained if we apply the TR-BDF2 method to the stiff source term in the Burgers-equation example. The use of this method eliminates the oscillations and unphysical states that were seen in Figure 17.8.

However, it still produces an incorrect solution in the stiff case, as seen in Figure 17.9(a). The wave now looks reasonable but is traveling at the wrong speed. It behaves fine if the grid is further refined, or equivalently if the value of τ is increased as shown in Figure 17.9(b). But the method still fails on underresolved grids in spite of the good behavior of the ODE solver.

This illustrates a difficulty often seen with stiff source terms. Similar behavior was observed by Colella, Majda & Roytburd [82] for detonation waves, and the problem has been discussed in many papers since. Even the simplest advection equation together with the source term (17.74) gives similar results. An analysis of this model problem, presented in LeVeque & Yee [293], carries over to the Burgers-equation example.



(a)

(b)

Fig. 17.9. Numerical solution to the Burgers equation with a stiff source term using the TR-BDF2 method for the source term: (a) the stiff case $\tau = 10^{-5}$; (b) the nonstiff case $\tau = 0.1$. [claw/book/chap17/stiffburgers]

Suppose

$$U_i^n = \begin{cases} 0 & \text{if } i < I, \\ 1 & \text{if } i \geq I, \end{cases}$$

as suggested by the result plotted in Figure 17.9(a). Then a high-resolution method for the Burgers equation will reduce to Godunov's method (since all slopes will be limited to 0), and we will obtain

$$U_i^* = \begin{cases} 0 & \text{if } i < I, \\ 1 - \Delta t/(2\,\Delta x) & \text{if } i = I, \\ 1 & \text{if } i > I, \end{cases}$$

since the jump propagates at speed 1/2. For the calculation shown in Figure 17.9(a), $U_I^* = 1 - \Delta t/(2\,\Delta x) = 0.65$. We now solve the stiff ODE, and since $U_I^* < \beta = 0.8$, the solution decays rapidly towards the equilibrium at 0, and so $U_I^{n+1} \approx 0$ after time Δt. In all other cells the source term is zero. We thus obtain

$$U_i^{n+1} \approx \begin{cases} 0 & \text{if } i < I + 1, \\ 1 & \text{if } i \geq I + 1, \end{cases}$$

and the wave has shifted over by one grid cell. The wave thus propagates at a speed of one grid cell per time step, which is a purely numerical artifact and not the correct physical speed.

Note that with a smaller time step we would have $U_I^* > \beta$, in which case $U_I^{n+1} \approx 1$, and so $U^{n+1} \approx U^n$. Now the numerical solution remains stationary (propagates with speed 0), again not the correct solution.

This difficulty arises from the fact that the numerical method for the homogeneous hyperbolic equation introduces numerical diffusion, leading to a smearing of the discontinuity or steep gradient that should be observed physically. The source term is then active over the entire region where the solution is smeared, leading to a larger contribution from this term than is physical. As we saw in Section 17.12, the speed at which the jump propagates is directly related to the strength of the source concentrated at the jump, so an incorrect net source term leads to incorrect propagation speeds.

This effect has been illustrated here with a fractional-step method. Other numerical methods, e.g., an unsplit method as in Section 17.2.1, can lead to similar difficulties. See [293] for an example. For some problems with stiff source terms it may be necessary to resolve the fastest scales (at least locally) in order to obtain good results. For a given value of τ the methods do converge as Δt, $\Delta x \to 0$. However, if τ is very small, then this may be impractical and we would prefer to capture the proper behavior on an underresolved grid. For some particular problems special methods have been developed that avoid the need for such finely resolved grids by calculating more accurately the correct source contribution; see for example [196] for one wave-propagation approach for a simple detonation model. See [19], [33], [40], [357], [348], [494] for some other possible approaches and further discussion of these numerical difficulties.

We should note that not all problems with stiff source terms lead to numerical difficulties on underresolved grids. For some problems the correct macroscopic behavior (in particular, correct propagation speeds) is observed even if the fast time scales are not well resolved.

To understand why, it is useful to consider the equation

$$q_t + f(q)_x = \psi(q) + \epsilon q_{xx}, \tag{17.89}$$

in which a diffusive or viscous term has been added in addition to the source term, which we assume is stiff with some fast time scale $\tau \ll 1$. Denote the solution to this equation by $q^{\tau,\epsilon}$. In practice there is typically physical dissipation present, but (as usual with the vanishing-viscosity approach) we assume this is very small, and we wish to find the limit $\lim_{\epsilon\to 0} q^{\tau,\epsilon}$ where τ is fixed at some physically correct value. On an underresolved grid we cannot hope to capture the detailed behavior on this fast time scale, and so we really seek to compute an approximation to

$$\lim_{\tau\to 0}\bigl(\lim_{\epsilon\to 0} q^{\tau,\epsilon}\bigr). \tag{17.90}$$

However, the numerical method will typically introduce dissipation $\epsilon > 0$ that depends on Δx, so that on an underresolved grid we can easily have $\epsilon \gg \tau$. As we refine the grid we are really approximating

$$\lim_{\epsilon\to 0}\bigl(\lim_{\tau\to 0} q^{\tau,\epsilon}\bigr), \tag{17.91}$$

at least as long as we remain on underresolved grids. It is only when we reach a grid that resolves the fast scale (which we don't want to do) that we would begin to approximate the correct limit (17.90). Pember [356] has conjectured that the problems leading to numerical difficulties on underresolved grids are precisely those for which the limits $\epsilon \to 0$ and $\tau \to 0$ do not commute, i.e., for which (17.90) and (17.91) give different results.

This is the case for the stiff Burgers-equation example of Section 17.15, for example. Adding a viscous term to (17.73) gives

$$u_t + \frac{1}{2}(u^2)_x = \frac{1}{\tau}u(1-u)(u-\beta) + \epsilon u_{xx}. \tag{17.92}$$

This equation has an exact traveling-wave solution that generalizes (17.80),

$$u(x,t) = \frac{1}{2}(1 + \tanh(\mu(x - st))), \tag{17.93}$$

where the values μ and s are given by

$$\mu = \frac{1}{\tau}\bigl(1 + \sqrt{1 + 8\epsilon/\tau}\bigr)^{-1}$$

and

$$s = \left(\frac{\beta}{2} - \frac{1}{4}\right)\bigl(1 + \sqrt{1 + 8\epsilon/\tau}\bigr) + \frac{1}{2}.$$

This solution can be found using results in [254], where a more general equation (with a cubic source term) is studied. For fixed τ, as $\epsilon \to 0$ we recover $\mu = 1/2\tau$ and $s = \beta$, so that (17.93) agrees with (17.80). On the other hand, if τ is smaller than ϵ, then this solution is quite different and the limits $\epsilon \to 0$ and $\tau \to 0$ do not commute.

In some physical problems adding additional dissipation does not substantially change the solution, and for such problems good results can often be obtained on underresolved grids. In the next section we study one class of problems for which this is true.

17.17 Relaxation Systems

In many physical problems there is an equilibrium relationship between the variables that is essentially maintained at all times. If the solution is perturbed away from this equilibrium, then it rapidly *relaxes* back towards the equilibrium. The problem considered in Section 17.15 has this flavor, but in that case there is a single variable u and two possible stable equilibria $u = 0$ and $u = 1$, leading to numerical difficulties. In this section we consider the situation where there are several variables and perhaps many different equilibrium states, but there is a unique equilibrium relationship between the variables.

A simple model problem is the system of two equations

$$\begin{aligned} u_t + v_x &= 0, \\ v_t + au_x &= \frac{f(u) - v}{\tau} \end{aligned} \tag{17.94}$$

for $0 < \tau \ll 1$, where $a > 0$ and $f(u)$ is a given function. The equilibrium states are those for which $v = f(u)$. If $v \neq f(u)$, then in the second equation the right-hand side dominates the term au_x, and v is rapidly driven back towards equilibrium (except perhaps in narrow reaction zones where u has a very steep gradient).

Note that if we simply assume $v \equiv f(u)$ in (17.94), then we can discard the second equation. Inserting this equilibrium assumption into the first equation, we then obtain the scalar conservation law

$$u_t + f(u)_x = 0. \tag{17.95}$$

Therefore we might expect solutions to the system (17.94) to be well modeled by the *reduced equation* (17.95) for small τ. In fact this is true, provided that a so-called *subcharacteristic condition* is satisfied. Observe that the homogeneous part of the system (17.94) is a linear hyperbolic system. The coefficient matrix has eigenvalues $\lambda^{1,2} = \pm\sqrt{a}$. Hence information can propagate no faster than speed $\sqrt{a}$, and adding the source term in (17.94) does not change this fact. On the other hand, the reduced equation (17.95) has characteristic speed $f'(u)$. If $|f'(u)| > \sqrt{a}$, then we cannot expect the solution of (17.94) to behave well in the limit $\tau \to 0$. The subcharacteristic condition for this problem is the requirement

$$-\sqrt{a} \leq f'(u) \leq \sqrt{a}. \tag{17.96}$$

Similar conditions hold for larger systems involving relaxation, and generally state that the characteristic speed of the reduced equation must fall within the range spanned by the characteristic speeds of the homogeneous part of the original system. This terminology was introduced by Liu [310], who studied more general relaxation systems in which au is replaced by a possibly nonlinear function $\sigma(u)$ in (17.94). Such problems arise in nonlinear elasticity (see Section 2.12.4), in which case $\sigma(u)$ represents the stress as a function of strain.

Many physical systems contain rapid relaxation processes that maintain an equilibrium, and often we solve the reduced equations based on this equilibrium. In fact the Euler equations of gas dynamics can be viewed as the reduced equations for more accurate models of the physics that include relaxation of vibrational or chemical nonequilibrium states that arise from intermolecular collisions. For many practical purposes the Euler equations are sufficient, though for some problems it is necessary to consider nonequilibrium flows; see for example [69], [70], [474], or [497]. A related example of the Euler equations relaxing towards isothermal flow is given in Section 17.17.3.

Relaxation systems are also used in modeling traffic flow. In Section 11.1 we derived a scalar conservation law by specifying the velocity of cars as a function of density $U(\rho)$. This assumes that drivers can react infinitely quickly to changes in density and are always driving at the resulting *equilibrium velocity*. In fact the velocity should relax towards this value at some rate depending on the drivers' reaction time. This can be modeled with a system of two equations, one for the density and a second for the velocity. Such models are often called *second-order models* in traffic flow (the scalar equation is the first-order model) and were first studied by Payne [354] and Whitham [486]. For more recent work, see for example [16], [99], [260], [296], [400], [498].

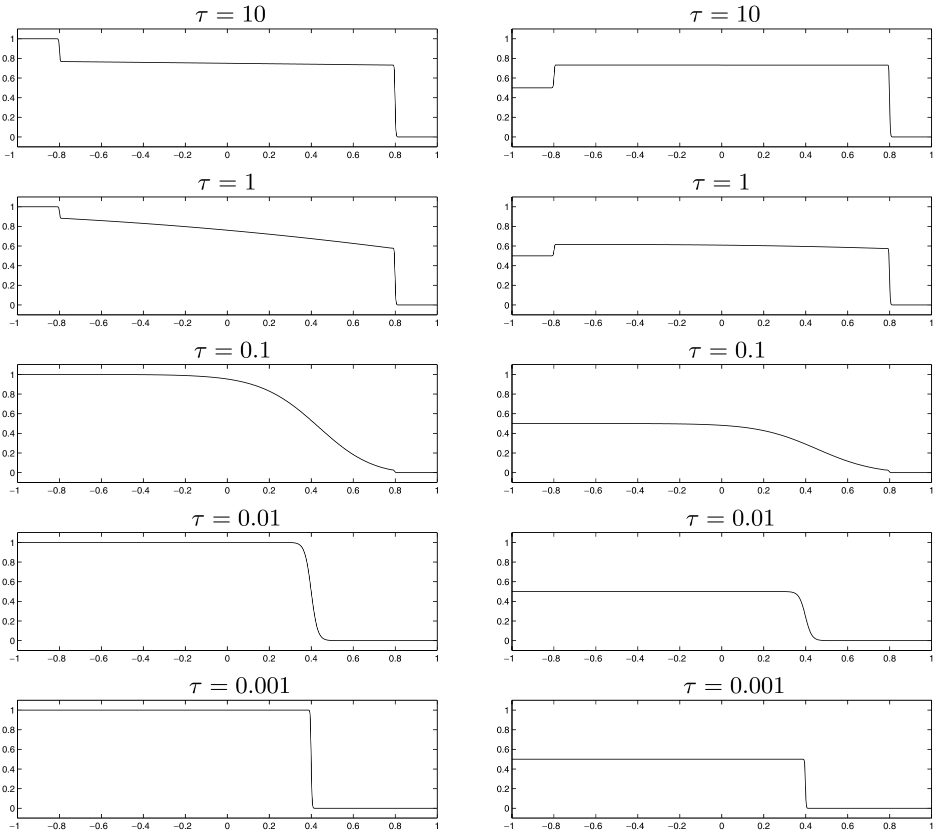
On the theoretical side there is considerable interest in the study of quite general relaxation systems and the conditions under which convergence of solutions occurs as $\tau \to 0$. See for example [64], [62], [305], [310], [336], [337], [491], [495]. Numerically it is useful to note that these relaxation problems can typically be solved on underresolved grids, as the next example shows. (This fact is the basis for the class of numerical *relaxation schemes* discussed in Section 17.18, in which stiff relaxation terms are intentionally introduced.)

Example 17.7. Figure 17.10 shows some solutions to the relaxation system (17.94) for $f(u) = \frac{1}{2}u^2$, $a = 1$, and the Riemann data

$$u_l = 1, \quad u_r = 0, \qquad v_l = f(u_l) = \frac{1}{2}, \qquad v_r = f(u_r) = 0. \tag{17.97}$$

The reduced equation is Burgers' equation, and we expect a shock traveling with speed $1/2$ in the limit as $\tau \to 0$, so it should be at $x = 0.4$ at the time $t = 0.8$ shown in all figures. Numerical solutions to the system (17.94) for various values of τ are shown, obtained using a fractional-step method. Solving the Riemann problem for the homogeneous linear system gives two waves with speeds ± 1. The state between these waves is not in equilibrium even if the Riemann data is, and the source term relaxes this state towards equilibrium. When τ is large (slow relaxation), the structure of the linear system is clearly visible, but as $\tau \to 0$ we observe convergence to the solution of Burgers' equation.

Note that for this problem the numerical fractional-step method behaves well in the limit $\tau \to 0$. The steep gradient near $x/t = 0.4$ (where the source term is active) is not well resolved on the grid for the smallest value of τ used in Figure 17.10, and yet the correct macroscopic behavior is observed. This is a reflection of the fact that if we add $\mathcal{O}(\epsilon)$ viscosity to the system (17.94), then the limits $\epsilon \to 0$ and $\tau \to 0$ commute, as described in Section 17.16. The basic reason for this is that the parameter τ also acts like a viscosity in the system (though this is not clear from the form of (17.94)), and so adding additional viscosity does not change the nature of the solution.


τ = 10

τ = 10

τ = 1

τ = 1

τ = 0.1

τ = 0.1

τ = 0.01

τ = 0.01

τ = 0.001

τ = 0.001

Fig. 17.10. Solution to the relaxation system (17.94) for five different values of τ, all shown at time $t = 0.8$. The left column shows u, and the right column shows v.

17.17.1 Chapman–Enskog Expansion

To see that τ is like a viscosity in the relaxation system (17.94), we use a so-called *Chapman–Enskog expansion*,

$$v(x,t) = f(u(x,t)) + \tau v_1(x,t) + \tau^2 v_2(x,t) + \cdots . \tag{17.98}$$

The form of this expansion is motivated by the fact that $v \to f(u)$ as $\tau \to 0$. Inserting this in the first equation of (17.94) gives

$$u_t + [f(u) + \tau v_1 + \tau^2 v_2 + \cdots]_x = 0,$$

or

$$u_t + f(u)_x = -\tau v_{1x} + \cdots . \tag{17.99}$$

To determine $v_1(x,t)$ we insert (17.98) in the second equation of (17.94), yielding

$$[f'(u)u_t + \tau v_{1t} + \tau^2 v_{2t} + \cdots] + au_x = -(v_1 + \tau v_2 + \cdots),$$

or by (17.99),

$$[f'(u)(-f(u)_x - \tau v_{1x} + \cdots) + \tau v_{1t} + \tau^2 v_{2t} + \cdots] + au_x = -(v_1 + \tau v_2 + \cdots).$$

Equating the $\mathcal{O}(1)$ terms for $\tau \ll 1$ gives

$$-f'(u)f(u)_x + au_x = -v_1$$

and hence

$$v_1 = -[a - f'(u)^2]u_x.$$

Using this in (17.99) gives

$$u_t + f(u)_x = \tau(\beta(u)u_x)_x + \mathcal{O}(\tau^2), \tag{17.100}$$

where

$$\beta(u) = a - [f'(u)]^2.$$

The equation (17.100) is a refined version of the reduced equation (17.95). For $\tau > 0$ we see that this parameter plays the role of a viscosity provided that $\beta(u) > 0$, which is true exactly when the subcharacteristic condition (17.96) is satisfied.

17.17.2 Violating the Subcharacteristic Condition

It is interesting to ask what happens if the subcharacteristic condition is violated. If $f(u) = bu$ is a linear function and $|b| > \sqrt{a}$, then the solution will blow up along the characteristic $\text{sgn}(b)\sqrt{a}$ as $\tau \to 0$. This case has been studied in [292]. Adding some viscosity to the system can stabilize the solution, since then the system is parabolic and allows arbitrary propagation speeds. If f is nonlinear, then the nonlinearity may also stabilize the solution. There are some physical problems where this is of interest, notably the case of *roll waves* in shallow water theory on a sloping surface [223], [486]. However, for the vast majority of physical problems involving relaxation the appropriate subcharacteristic condition is satisfied.

17.17.3 Thermal Relaxation and Isothermal Flow

As an example of a relaxation system we consider the Euler equations with a relaxation term driving the temperature towards a constant value. The equations of isothermal flow were introduced in Section 14.6, for gas in a one-dimensional tube surrounded by a bath at constant temperature. We assume that heat flows in or out of the bath instantaneously, so that a constant temperature is maintained in the gas (so heat is extracted from the gas just behind a shock wave, and flows into the gas in a rarefaction wave).

This is obviously not a perfect model of the physical situation. A better model would be the relaxation system

$$\begin{bmatrix} \rho \\ \rho u \\ E \end{bmatrix}_t + \begin{bmatrix} \rho u \\ \rho u^2 + p \\ (E+p)u \end{bmatrix}_x = \begin{bmatrix} 0 \\ 0 \\ -[E - \bar{E}(\rho, \rho u)]/\tau \end{bmatrix} \tag{17.101}$$

together with an appropriate equation of state for the gas, e.g., (14.23) for a polytropic ideal gas. Here $\bar{E}(\rho, \rho u)$ is the energy in the gas that results if we bring T to the bath temperature $\bar{T}$ without changing the density or momentum. According to the ideal gas law (14.9), we then have $p = R\rho\bar{T} = a^2\rho$, where $a = \sqrt{R\bar{T}}$ is the isothermal sound speed. Using this in the equation of state (14.23) gives

$$\bar{E}(\rho, \rho u) = \frac{a^2\rho}{\gamma - 1} + \frac{1}{2}\rho u^2. \tag{17.102}$$

The quantity $\tau > 0$ is the time scale over which the energy $E - \bar{E}$ flows into the tube, the reciprocal of the *relaxation rate*. The isothermal equations (14.34),

$$\begin{bmatrix} \rho \\ \rho u \end{bmatrix}_t + \begin{bmatrix} \rho u \\ \rho u^2 + a^2\rho \end{bmatrix}_x = 0, \tag{17.103}$$

result from letting $\tau \to 0$. These are the reduced equations corresponding to the relaxation system (17.101).

Note that the subcharacteristic condition is satisfied for this system provided that $a \le c$, where a is the isothermal sound speed and $c = \sqrt{\gamma p/\rho}$ is the sound speed for the full polytropic Euler equations. Near equilibrium we have $p \approx (\gamma - 1)(\bar{E} - \frac{1}{2}\rho u^2) = R\bar{T}\rho$, and so $c = \sqrt{\gamma R\bar{T}}$. Since $\gamma > 1$ we have $c > a$.

The system (17.101) has a structure similar to (17.94). We can divide the variables $(\rho, \rho u, E)$ into *reduced variables* $(\rho, \rho u)$ and the *relaxation variable* E, which relaxes quickly to an equilibrium state $\bar{E}$, a unique value for any given $(\rho, \rho u)$. The reduced equation is obtained by assuming $E \equiv \bar{E}(\rho, \rho u)$ and eliminating the relaxation variable. The system (17.94) has the same structure, with u being the reduced variable and v the relaxation variable. This structure is important in that it allows us to perform a Chapman–Enskog expansion as in Section 17.17.1 to show that the relaxation time τ plays the role of a viscosity in the reduced equation. This in turn suggests that numerical methods will be successful on these relaxation systems even on underresolved grids.

In particular, it is important here that there is a unique equilibrium state $\bar{E}$ corresponding to any given values of the reduced variables $(\rho, \rho u)$. The relaxation variable E converges to this value regardless of the initial value of E. By contrast, in the example of Burgers' equation with a stiff source term, (17.73), there is no separation into reduced and relaxation variables, since this is a scalar equation and u plays both roles. There are two possible equilibrium states $u = 0$ and $u = 1$, and which one we relax towards depends on the initial value of u. This means that a smearing of u can lead to the calculation of an incorrect equilibrium state, resulting in incorrect wave speeds, as was illustrated in Section 17.16. Note also that the example of Section 17.16 suggests that τ does not play the role of a viscosity in the stiff Burgers example. Adding viscosity to a rarefaction wave should leave it essentially unchanged, whereas the stiff source term of (17.73) converts the rarefaction wave into a thin reaction zone that approaches a discontinuity as $\tau \to 0$.

17.18 Relaxation Schemes

In the example of the previous section, the relaxation system is the "correct" physical model and the reduced system is an approximation valid in the limit $\tau \to 0$. In practice one might want to use the reduced system (e.g., the isothermal equations) instead of solving the more complicated relaxation system, which involves stiff source terms as well as being a larger system of equations. Similarly, the scalar conservation law (17.95) can be viewed as an approximation to the relaxation system (17.94).

In some situations we may wish to turn this viewpoint around and instead view a relaxation system such as (17.94) as an approximation to the conservation law (17.95). Suppose we wish to solve the nonlinear conservation law $u_t + f(u)_x = 0$ numerically but for some reason don't wish to write a nonlinear Riemann solver for $f(u)$. Then we can artificially introduce a relaxation variable v and instead solve the system (17.94) for some a and τ, using a fractional-step method. This only requires solving a *linear* hyperbolic system with simple characteristic structure in each time step. For a scalar problem this may be pointless, since the nonlinear Riemann problem is typically easy to solve, but for a nonlinear system of m conservation laws the same idea can be applied, by using the relaxation system

$$\begin{aligned} u_t + v_x &= 0, \\ v_t + Au_x &= \frac{f(u) - v}{\tau}, \end{aligned} \tag{17.104}$$

where $u,\ v \in \mathbb{R}^m$ and A is some $m \times m$ matrix. The original nonlinear system of m equations is converted into a linear system of $2m$ equations with the nonlinearity concentrated in the source terms.

The coefficient matrix of (17.104) has the form

$$B = \begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix}, \tag{17.105}$$

with eigenvalues $\pm\sqrt{\lambda}$, where λ is an eigenvalue of A. In order for the relaxation system to be strictly hyperbolic, we require $\lambda > 0$ for each eigenvalue of A, i.e., A must be positive definite. We also need the subcharacteristic condition to hold, which requires that the eigenvalues of the Jacobian matrix $f'(u)$ should always lie within the range of eigenvalues of B, i.e., within the interval $\pm \max \sqrt{\lambda^p}$. This is a generalization of the subcharacteristic condition (17.96) for the case $m = 1$.

This numerical approach was introduced by Jin and Xin [225]. Methods of this type are generally called *relaxation schemes*, and have since been extensively studied, e.g., [12], [60], [64], [164], [222], [226], [233], [261], [306], [439]. Jin and Xin present some theory and numerical results for the Euler equations, taking A to be a diagonal matrix so that the Riemann problem for (17.104) is easy to solve. The choice of elements of A must be based on estimates of the minimum and maximum wave speed that will arise in the problem so that the subcharacteristic condition is satisfied. In [225] the stiff source term is solved using an implicit Runge–Kutta method with a small value of τ. In practice it seems to work as well to simply consider the limit $\tau \to 0$, called the *relaxed scheme* in [225]. We can then implement a single step of the relaxation scheme using a fractional-step method of the following form:

1. U^n, V^n are used as data for the homogeneous version of (17.104). Solving this linear hyperbolic system over time Δt produces values U^*, V^*.
2. These values are used as data for the system with only the source terms, so $u_t = 0$ and $v_t = [f(u) - v]/\tau$. The solution of this in the limit $\tau \to 0$ has u constant and $v \to f(u)$, so we simply set

$$\begin{aligned} U^{n+1} &= U^*, \\ V^{n+1} &= f(U^{n+1}). \end{aligned} \tag{17.106}$$

With this approach we do not need to choose a specific value of τ or solve the ODE; we simply solve the linear hyperbolic system and then reset V to $f(U)$. Note that the success of relaxation schemes depends on the fact that, for relaxation systems, the stiff source terms do not cause numerical difficulties on underresolved grids.

The advantage of a relaxation scheme is that it avoids the need for a Riemann solver for the nonlinear flux function $f(u)$, since the hyperbolic system to be solved now involves the linear coefficient matrix (17.104) instead. The relaxation scheme can alternatively be viewed as a particular way of defining an approximate Riemann solver for the nonlinear problem, as shown in [288], which results in some close connections with other approximate Riemann solvers introduced in Section 15.3. A particularly simple choice of A leads to the Lax–Friedrichs method (4.20); see Exercise 17.9.

Exercises

17.1. Suppose $\beta(x)$ varies with x in the problem of Section 17.2. Derive an unsplit second-order accurate method for this problem.

17.2. Compute the $\mathcal{O}(\Delta t^3)$ term in the splitting error for the Strang splitting (17.35).

17.3. Determine the splitting error for the Godunov splitting applied to the system (17.4).

17.4. Suppose we wish to numerically approximate the steady-state solution to (17.7) for the boundary-value problem with $q(0, t) = C$. The exact solution is given by (17.44).

(a) Consider the unsplit method (17.9), and suppose we have reached a numerical steady state so that $Q_i^{n+1} = Q_i^n$ for all i. Show that this numerical steady-state solution satisfies

$$Q_i = \frac{Q_{i-1}}{1 + \beta\,\Delta x/\bar{u}}.$$

Hence the numerical steady-state solution is accurate to $\mathcal{O}(\Delta x)$ and is independent of Δt.

(b) Now consider the fractional-step method (17.21), and show that the numerical steady-state solution satisfies

$$Q_i = \frac{Q_{i-1}}{1 + (\beta\,\Delta x/\bar{u})(1 - \beta\,\Delta t)}.$$

Again this agrees with the true steady-state solution to $\mathcal{O}(\Delta x) = \mathcal{O}(\Delta t)$, but now the numerical steady state obtained depends on the time step Δt.

17.5. Apply the approach of Section 17.14 to the advection–reaction problem (17.7), with $\Psi_{i-1/2} = -\frac{\beta}{2}(Q_{i-1} + Q_i)$. Use the resulting fluctuations in (4.43) to determine a first-order accurate unsplit method for this equation. How does this method compare to (17.9)? Is the resulting numerical steady state independent of Δt?

17.6. Determine the structure of the on-ramp Riemann problem of Section 17.13. In particular, show that a shock forms whenever $D > (1 - 2q_l)^2/4$.

17.7. Solve the relaxation system (17.101) numerically using CLAWPACK and the Godunov splitting for the source terms. Compare the results on a shock-tube problem with results obtained by solving the isothermal equations. Try various values of τ and both resolved and underresolved grids.

17.8. Determine the eigenvectors of the matrix B in (17.105) in terms of the eigenvectors of A.

17.9. Take $A = (\Delta x/\Delta t)^2 I$ in the relaxation system (17.104). Show that the relaxed scheme described at the end of Section 17.18 then reduces to the Lax–Friedrichs method (4.20). How does the subcharacteristic condition relate to the CFL condition in this case?