Numerical Analysis Project
Manuscript NA-81-16

October, 1981

NUMERICAL METHODS BASED ON ADDITIVE SPLITTINGS FOR HYPERBOLIC PARTIAL DIFFERENTIAL EQUATIONS

b y

Randall J. LeVeque

Joseph Oliger

Numerical Analysis Project
Computer Science Department
Stanford University
Stanford, California 94305

Numerical Methods based on Additive Splittings for Hyperbolic Partial Differential Equations

Randall J. LeVeque*
Joseph Oliger*

Abstract. We derive and analyze several methods for systems of hyperbolic equations with wide ranges of signal speeds. These techniques are also useful for problems whose coefficients have large mean values about which they oscillate with small amplitude. Our methods are based on additive splittings of the operators into components that can be approximated independently on the different time scales, some of which are sometimes treated exactly. The efficiency of the splitting methods is seen to depend on the error incurred in splitting the exact solution operator. This is analyzed and a technique is discussed for reducing this error through a simple change of variables. A procedure for generating the appropriate boundary data for the intermediate solutions is also presented.

*Department of Computer Science, Stanford University, Stanford, CA 94305.

Both authors have been supported in part for this work by the Office of Naval Research under contract N00014-75-C-1132 and the National Science Foundation under grant MCS77-02082. The first author has also been supported by National Science Foundation and Hertz Foundation graduate fellowships.

1. Introduction.

Splitting methods for time-dependent partial differential equations have been most frequently studied in the context of spatial splittings, as in the approximate factorization techniques for efficiently implementing implicit algorithms in more than one space dimension[6], [11], [13]. Some attention has also been given to splitting or fractional step methods for problems where the differential operator is split up into pieces corresponding to different physical processes which are most naturally handled by different techniques. This has been done, for example, with convection-diffusion and the Navier-Stokes equations[1], [4], [5].

More generally, a splitting method may be useful any time one is faced with a problem

$$u_t = Au \tag{1.1}$$

where A is some differential operator of the form

$$A = A_1 + A_2$$

such that the problems

$$u_t = A_1 u \tag{1.2a}$$

and

$$u_t = A_2 u \tag{1.2b}$$

are each easier to solve than the original problem. By alternating between solving (1.2a) and (1.2b) we hope to compute a satisfactory solution to (1.1).

In this paper we consider such methods applied to a one-dimensional quasilinear hyperbolic system

$$u_t = A(x, t, u)u_x \tag{1.3}$$

where A is an $n \times n$ matrix with real eigenvalues. We assume that A is of the form

$$A = A_f + A_s. \tag{1.4}$$

In our notation, "f" and "s" stand for "fast" and "slow" respectively, which reflects a common situation in which the solution contains waves traveling at quite different wave speeds. If A is constant then the solution operator for the problem (1.3) on a single tirnestep of size k is $\exp(kA\partial_x)$, that is to say

$$u(x, t + k) = \exp(kA\partial_x)u(x, t). \tag{1.5}$$

For nonconstant A the solution operator is more complicated. Our analysis will be concerned mostly with the constant coefficient case, so we will use the notation of (1.5) throughout. The ideas generalize easily, but are most intuitively seen in terms of exponentials.

The additive splitting (1.4) comes into play when the solution operator $\exp(kA\partial_x)$ is approximated by the product of the solution operators for the subproblems

$$u_t = A_f u_x \tag{1.6a}$$

and

$$u_t = A_s u_x. \tag{1.6b}$$

We replace (1.5) by

$$u(x, t + k) \approx \exp(kA_f\partial_x)\exp(kA_s\partial_x)u(x, t).$$

An approximation lo $u(x,\ t\ +$ k) is thus obtained by first solving (1.6b) with $u(x,\ t)$ as initial data, and using the resulting solution as initial data for (I .6a). If A_f anti A, commute, this splitting is exact. When they do not commute, we have introduced an error which is $O(k^2)$. As noted by Strang[16], this error can be reduced to $O(k^3)$ by USC of the splitting

$$\exp(k(A_f + A_s)\partial_x) \approx \exp(\tfrac{1}{2}kA_f\partial_x)\exp(kA_s\partial_x)\exp(\tfrac{1}{2}kA_f\partial_x). \tag{1.7}$$

Analogous results hold for the corresponding splittings with variable coefficients. Computations confirm that the global error is also improved (from O(k) to $O(k^2)$) by the use of this splitting.

The numerical approximations to the solution operators $\exp(kA_s\partial_x)$ and $\exp(\frac{1}{2}kA_f\partial_x)$ will be denoted by $Q_s(k)$ and $Q_f(k/2)$ respectively. The numerical method based on the Strang splitting (1.7) is then

$$U_m^{n+1} = Q_f(k/2)Q_s(k)Q_f(k/2)U_m^n \tag{1.8}$$

where U_m^n is the numerical approximation to $u(mh,$ nk) on a grid with Ax = h and At = k. When splitting a multidimensional problem into one-dimensional subproblems, this sort of a splitting gives rise to the so-called ***locally one dimensional (LOD) method,*** a spatially split scheme. In the present context we will refer to (1.8) as the ***time-split method.***

In practice U^{n+1} can be computed via the sequence

$$\begin{aligned} U_m^* &= Q_f(k/2)U_m^n \\ U_m^{**} &= Q_s(k)U_m^* \\ U_m^{n+1} &= Q_f(k/2)U_m^{**} \end{aligned} \tag{1.9}$$

'although it should be noted that when several steps of (1.8) are applied successively the adjacent $Q_f(k/2)$ operators can be combined into $Q_f(k)$, and the half-step operators need only be applied at the beginning and immediately before printout, i.e.

$$U_m^n = Q_f(k/2)Q_s(k)Q_f(k)\cdots Q_f(k)Q_s(k)Q_f(k/2)U_m^0.$$

There are several situations in which the use of the time split method may lead to a more efficient solution of the original problem. We will mention three such cases here. Our analysis will be mostly concerned with the first and last of these.

Problem 1: Suppose the solution to (1.3) contains both fast waves and slow waves, i.e. the eigenvalues μ_i of A satisfy

$$|\mu_1| \le |\mu_2| \le \cdots \le |\mu_p| \ll |\mu_{p+1}| \le \cdots \le |\mu_n|.$$

Assume also that there are relatively few elements of A which contribute to the fast waves. We can take advantage of this structure-by splitting the operator into slow and fast parts and using small time steps only on the fast part. That is, we can choose k so that $\exp(kA_s\partial_x)$ can be adcquatly represented by a single step of some finite difference scheme and then approximate $\exp(\frac{1}{2}kA_f\partial_x)$ by several steps of a difference scheme with a smaller timestep. Similarly, we can handle more than two clusters of wave speeds by means of further splittings.

Such a splitting method requires less work than using small timesteps on the full unsplit problem, and will thus be more efficient provided the accuracy is not too adversely affected by the error in the splitting; We will see that this depends very much on the problem at hand. In cases where the splitting error is small, the time-split method actually may be more accurate, since we

will be able to use nearly optimal mesh ratios for each cluster to minimize the truncation error and improve other characteristics of the method, such as its dissipative behavior.

Similar splittings have been considered by Engquist, Gustafsson and Vreeburg[4] for this type of problem. However, the **splitting** in their problem involved little interaction between the different time scales, so that many of the problems we shall encounter were not present.

Example 1.1. Consider a block triangular system of the form

$$A = \begin{bmatrix} \frac{1}{\epsilon}A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \tag{1.10}$$

with $\|A_{11}\| \approx \|A_{22}\| \approx 1$ and $\epsilon \ll 1$. It is reasonable to take

$$A_f = \begin{bmatrix} \frac{1}{\epsilon}A_{11} & 0 \\ 0 & 0 \end{bmatrix}, \quad A_s = \begin{bmatrix} 0 & A_{12} \\ 0 & A_{22} \end{bmatrix}. \tag{1.11}$$

For this problem the effectiveness of the split method depends greatly on the coupling A_{12} between the different time scales. This is analyzed in section 2 where we also present a simple procedure for changing variables to reduce the coupling.

Problem 2: Consider the same situation as in Problem 1, but where the fast waves are known to be absent from the physical solution of interest. Recently Kreiss[9], [IO] and Browning, Kasahara, and Kreiss[2] have considered some new approaches for this problem which rely upon properly preparing the data so that the fast wave components are eliminated. These can be considered as projection techniques. Majda[12] has considered using filters to suppress the fast waves in the same context.

In this case the true solution should satisfy

$$\exp(kA\partial_x)u(x,t) = \exp(kA_s\partial_x)u(x,t)$$

providing the splitting between fast and slow scales is done correctly. For variable coefficient problems it will not be possible to have the correct splitting at all times and the operator A_f cannot be dropped entirely. However, we can consider using the time-split method (t.8) with a less accurate scheme for $Q_f(k/2)$ than is used for $Q_s(k)$, perhaps by using the-same. timestep for both with a larger spatial step for $Q_f(k/2)$. In such a manner it may again be possible to obtain the same accuracy more efficiently. Turkel and Zwas[18] have considered a method for this problem which is similar in spirit.

Problem 3: Suppose that the coefficients in the problem (1.3) have large mean values about which they oscillate with small amplitude. In this case it may be possible to split out a constant coefficient problem which can be solved exactly, leaving behind the small perturbations for A_s. Then (1.8) can be used with some large timestep approximation for $Q_s(k)$ while $Q_f(k/2) = \exp(\frac{1}{2}kA_f\partial_x)$ exactly. This is clearly more efficient than using small timesteps on the unsplit problem. Moreover, since the dominant part of the operator is being handled exactly, great increases in accuracy are also possible.

Example 1.2. The simplest example is the scalar problem

$$u_t = (1 + \alpha(x))u_x \tag{1.12}$$

where $|\alpha(x)| << 1$ and we use the splitting'

$$A_f = 1, \qquad A_s = a(x).$$

Take $k = ph$ for some integer p. The operator $\exp(\frac{1}{2}kA_f\partial_x)$ in known exactly: $\exp(\frac{1}{2}kA_f\partial_x)u(x, t) = u(x + \frac{1}{2}ph, t)$. If Lax-Wendroff is used for the remaining subproblem $u_t = \alpha(x)u_x$ then the method (1.8) cnn be written as a single step method

$$U_m^{n+1} = U_{m+p}^n + \tfrac{1}{2}p(U_{m+p+1}^n - U_{m+p-1}^n) + \tfrac{1}{4}p^2\alpha(\hat{x}_m)\big((\alpha(\hat{x}_m + h) + \alpha(\hat{x}_m))(U_{m+p+1}^n - U_{m+p}^n) - (\alpha(\hat{x}_m) + \alpha(\hat{x}_m - h))(U_{m+p}^n - U_{m+p-1}^n)\big)$$

where $\hat{x}_m = x_m + \frac{1}{2}ph$. Notice that even though this is a scalar problem, the operators ∂_x and $\alpha(x)\partial_x$ do not commute and so the Strang spliting must be used.

The shallow water equations provide a more interesting example of Problem 3. These are discussed in section 8.

General considerations. The effectiveness of the time-split method depends on the error in the splitting (1.7). If this is exact then the splitting method is clearly more efficient For the types of problems we are considering. On the other hand, if the splitting error dominates, it may be necessary to reduce the timestep considerably, eliminating the possible benefits of the splitting. In the next section we derive an explicit expression for the splitting error and indicate how to determine whether a splitting method is useful on a given problem. We will show that the equations sometimes need to be transformed to reduce the linkage between fast and slow modes in order to achieve the desired accuracy.

The Q operators in the time-split method can consist of one or more steps of any explicit or implicit scheme using two time levels. It is not immediately clear how a scheme using more than two time levels (such as leapfrog) could be used. For suppose we want to use leapfrog as $Q,(k)$ to'go -from U^* to U^{**} in (1.9). Then we would need-some approximation to $\exp(-kA_s\partial_x)U^*$ (which is *not* U^n) to play the role of U^* at the previous time level. As a first step towards incorporating multi-level schemes into the splitting framework, section 4 introduces a different type of split scheme which does use leapfrog for $Q,(k)$. This method is based upon approximations to the variation of parameters formula, or Duhamel's principal, and will be called the Leapfrog Duhamel method. Similar ideas have been used for ordinary differential equations by Certaine[3]. The accuracy and stability of the Leapfrog Duhamel method is considered in sections 5 and 6.

The initial boundary value problem is considered in section 7. In most cases boundary data will have to be supplied for the intermediate solutions U^* and U^{**} in (1.9). We consider the problem of approximating the correct boundary data in terms of the given boundary conditions.

Some further examples of splittings and computational results are-presented in section 8.

2. Accuracy of the time-split method.

In this section we consider discretizations of the approximate splitting

$$u(x, t + k) \approx \exp(\tfrac{1}{2}kA_f\partial_x)\exp(kA_s\partial_x)\exp(\tfrac{1}{2}kA_f\partial_x)\,u(x, t) \tag{2.1}$$

for the solution of $u_t = Au, \equiv (A_f + A_s)u_x$ with $\|A_s\| \ll \|A_f\|$. Up until section 7 we will deal only with the Cauchy problem, where $-\infty < x < \infty$. Of course these results also hold for a strip problem with periodic boundary conditions, e.g., $0 \le x \le 1$ and $u(0, t) = u(1, t)$. We will assume that A_f and A_s are constant matrices but our approach carries over for more general problems if the exponentials in (2.1) are replaced by the appropriate solution operators. For example, the splitting error for the problem (1.12) is given in example 8.2 of section 8.

If A_f and A_s commute then the splitting (2.1) is exact. Otherwise we define the *splitting error*

operator $E_{\text{split}}(k)$ by

$$\begin{aligned} E_{\text{split}}(k) &= \exp(\tfrac{1}{2}kA_f\partial_x)\exp(kA_s\partial_x)\exp(\tfrac{1}{2}kA_f\partial_x) - \exp(k(A_f + A_s)\partial_x) \\ &= -\tfrac{1}{6}k^3(\tfrac{1}{4}A_f^2A_s - \tfrac{1}{2}A_fA_sA_f + \tfrac{1}{4}A_sA_f^2 \\ &\quad - \tfrac{1}{2}A_s^2A_f + A_sA_fA_s - \tfrac{1}{2}A_fA_s^2)\partial_x^3 + O(k^4). \end{aligned} \tag{2.2}$$

The ***local truncation error operators*** for the approximate solution operators $Q_f(k/2)$ and $Q,(k)$ are defined by

$$\begin{aligned} E_f(k/2) &= Q_f(k/2) - \exp(\tfrac{1}{2}kA_f\partial_x) \\ E_s(k) &= Q_s(k) - \exp(kA_s\partial_x). \end{aligned}$$

Note that for a second order scheme such as Lax-Wcndroff these are $O(k^3)$. We can now compute the truncation error operator for the split scheme. The numerical solution operator is

$$\begin{aligned} Q_f(k/2)Q_s(k)Q_f(k/2) &= (\exp(\tfrac{1}{2}kA_f\partial_x) + E_f(k/2))(\exp(kA_s\partial_x) + E,(k)) \\ &\quad \times (\exp(\tfrac{1}{2}kA_f\partial_x) + E_f(k/2)) \\ &= \exp(\tfrac{1}{2}kA_f\partial_x)\exp(kA_s\partial_x)\exp(\tfrac{1}{2}kA_f\partial_x) \\ &\quad + E,(k) + 2E_f(k/2) + O(k^4) \\ &= \exp(k(A_f + A_s)\partial_x) + E_{\text{split}}(k) \\ &\quad + E,(k) + 2E_f(k/2) + O(k^4). \end{aligned}$$

-The truncation error operator for the time-split method (TSM) is thus

$$\begin{aligned} E^{TSM}(k) &= Q_f(k/2)Q_s(k)Q_f(k/2) - \exp(k(A_f + A_s)\partial_x) \\ &= E_{\text{split}}(k) + E_s(k) + 2E_f(k/2) + O(k^4). \end{aligned} \tag{2.3}$$

For a given problem this error can be computed directly and used to assess the efficiency of the time-split method relative to an unsplit method.

In order to illustrate some of the propertics of this method and the effect the splitting error $E_{\text{split}}(k)$ has on its utility, we restrict our attention to the case where $Q,(k)$ consists of a single step of Lax-Wendroff. For convenience we use $LW(A, k)$ to denote the Lax-Wendroff operator,

$$LW(A, k) = I + kAD_0 + \tfrac{1}{2}k^2A^2D_+D_-$$

where D_0, $D+$ and D_- are the standard centered, forward, and backward difference operators, respectively. We thus have

$$Q_s(k) = LW(A_s, k). \tag{2.4a}$$

For $Q_f(k/2)$ we consider both

$$Q_f(k/2) = \exp(\tfrac{1}{2}kA_f\partial_x) \tag{2.4b}$$

and

$$Q_f(k/2) = (LW(A_f, k/m))^{m/2} \tag{2.4c}$$

for some even integer m. The situation (2.4b) occurs when the solution operator $\exp(\tfrac{1}{2}kA_f\partial_x)$ is known exactly, as in Problem 3. In (2.4c) $Q_f(k/2)$ consists of $m/2$ steps of Lax-Wendroff with timestep k/m. This might be appropriate when solving Problem 1, for example.

The standard error analysis for Lax-Wendroff shows that for (2.4a) we have

$$E_s(\mathsf{k}) = E_s^{LW}(\mathsf{k}) = -\tfrac{1}{6}(k^3A_s^3 - kh^2A_s)\partial_x^3 + O(k^4). \tag{2.5a}$$

When (2.4b) is used there is no error on the fast, scale and

$$E_f(k/2) = E_f^{\exp}(k/2) = 0. \tag{2.5b}$$

Otherwise, when (2.4c) is used,

$$\begin{aligned} Q_f(k/2) &= (LW(A_f, k/m))^{m/2} \\ &= \left(\exp(\tfrac{k}{m}A_f\partial_x) - \tfrac{1}{6}(\tfrac{k^3}{m^3}A_f^3 - \tfrac{k}{m}h^2A_f)\partial_x^3\right)^{m/2} \\ &= \exp(\tfrac{1}{2}kA_f\partial_x) - \tfrac{1}{6}\left(\tfrac{k^3}{2m^2}A_f^3 - \tfrac{k}{2}h^2A_f\right)\partial_x^3 + O(k^4) \end{aligned}$$

and so

$$E_f(k/2) = E_f^{LW}(k/2) = -\tfrac{1}{12}(\tfrac{k^3}{m^2}A_f^3 - kh^2A_f)\partial_x^3 + O(k^4). \tag{2.5c}$$

In this case WC must choose an appropriate value of m, the number of small timesteps used within each large timestep. For fixed h, the error $E_f^{LW}(k/2)$ does not approach zero as $m \to \infty$. From (2.5c) it seems unreasonable to take m any larger than a value for which $\|\frac{k^3}{m^2}A_f^3\| \approx \|kh^2A_f\|$. This suggests taking

$$m \approx \frac{k}{h}\|A_f\| \tag{2.6}$$

The proper choice of m may also be influenced by stability requirements. Determining the stability of the operator $Q_f(k/2)Q_s(k)Q_f(k/2)$ is in general a difficult problem, which will be considered in some detail in section 5. It will be shown there that for some problems the product operator is stable provided $Q_f(k/2)$ and $Q_s(k)$ are each stable independently. It is well known that for Lax-Wendroff the stability condition on $Q_f(k/2)$ is $\rho(A_f)k/mh \leq 1$, i.e., $m \geq \rho(A_f)k/h$. The m given in (2.6) is consistent with this requirement. Also note that for $k/h \approx 1/\|A_s\|$, (2.6) becomes $m \approx \|A_f\|/\|A_s\|$.

When the splitting error $E_{\text{split}}(k)$ is negligible compared to the other terms in (2.3), the truncation error for the split scheme becomes

$$\begin{aligned} E^{TSM}(k) &= E_s(k) + 2E_f(k/2) + O(k^4) \\ &= -\tfrac{1}{6}(k^3(A_s^3 + \tfrac{1}{m^2}A_f^3) - kh^2(A_s + A_f))\partial_x^3 + O(k^4). \end{aligned}$$

This error is roughly the same as we would obtain using $(LW(A, k/m))^m$, i.e., Lax-Wendroff with small steps on the unsplit problem. The truncation error for the unsplit method can be derived in the same manner as (2.5c) to obtain

$$\begin{aligned} E^{LW}(k) &= -\tfrac{1}{6}(\tfrac{k^3}{m^2}(A_f + A_s)^3 - kh^2(A_f + A_s))\partial_x^3 + O(k^4) \\ &\quad - \tfrac{1}{6}(\tfrac{k^3}{m^2}A_f^3 - kh^2(A_f + A_s))\partial_x^3. \end{aligned} \tag{2.7}$$

Thus we do almost as well by taking large steps with A_s and small steps with A_f as we would by taking small steps on the unsplit problem. This can lead to considerable savings. If $Q_f(k/2) =$

$\exp(\frac{1}{2}kA_f\partial_x)$ the results are even more striking. Now the error (2.3) is simply

$$E^{TSM}(k) = E,(k) + O(k^4) = -\tfrac{1}{6}(k^3A_s^3 - kh^2A_s)\partial_x^3 + O(k^4).$$

Comparing this to (2.7) shows that the split scheme is considerably more accurate. It also requires less work, since now nothing is computed using small steps.

The results of the last paragraph are all based on the assumption that $E_{\text{split}}(k)$ is negligible. In practice $E_{\text{split}}(k)$ may easily dominate the discretization error E,(k) + $2E_f(k/2)$. in this case the split scheme is less accurate than Lax-Wendroff with timestep k/m. Nonetheless the split scheme may be preferable. It may be possible to use the split scheme with smaller k and h to obtain better accuracy while still requiring less work than the unsplit scheme. The proper quantity for comparison is the work required to obtain a given accuracy. This can be estimated and compared for various methods as we now do. Under some mild assumptions, we will see that the methods (2.4a,b) and (2.4a,c) are always more efficient than the unsplit scheme (providing we choose k/h properly).

Work comparisons. We will compute expressions for the work required to obtain a solution at time $t = 1$ with error at most τ. All of the bounds below are rough order of magnitude bounds which are sufficient for our present purpose. Suppose that

$$\|A\| \approx \|A_f\| = a, \qquad \|A_s\| = b$$

where $b/a = \epsilon << 1$. Also suppose that $\|u_{xxx}\| \approx 1$. This is for convenience only, since it removes one common factor from all of the bounds below.

We will first analyze the unsplit Lax-Wendroff method $LW(A, k)$. Suppose that W is the work required to compute $LW(A, k)U_m^n$ at a single point x_m. Then the work required to advance the solution on a unit z-interval by one unit of time is $W/kh = \lambda W/k^2$ if $k = Xh$. The error committed in one time unit using the unsplit method is bounded by

$$\begin{aligned}\|((LW(A,k))^{1/k} - \exp(A\partial_x))u\| &\\ \le \frac{1}{k}(\tfrac{1}{6}(k^3\|A\|^3 + kh^2\|A\|) + O(k^4)) &\\ \le \tfrac{1}{6}k^2(a^3 + a/\lambda^2) + O(k^3).&\end{aligned}$$

Since we require an error $\approx \tau$, we set

$$\tfrac{1}{6}k^2(a^3 + a/\lambda^2) = \tau$$

giving

$$k^2 = \frac{6\tau}{a(a^2 + 1/\lambda^2)}.$$

Thus $w(\tau; \lambda)$, the work required to achieve a given accuracy τ using Lax-Wendroff with stepsize ratio λ, is given by

$$\begin{aligned}w(\tau;\lambda) &= \frac{\lambda W}{k^2} \\ &= (\lambda a + 1/\lambda a)\frac{a^2W}{6\tau}.\end{aligned}$$

We have not yet specified λ. Choosing λ to minimize $w(\tau;\lambda)$ gives $\lambda = 1/a$ and the minimum work $w(\tau)$ is

$$W(T) = \frac{a^2 W}{3\tau} \qquad \text{for unsplit Lax-Wendroff.} \tag{2.8}$$

Now consider the split method (2.4a,b). Let W_s be the work required to apply Lax-Wendroff on the slow scale and $W_f^{\exp}$ the work required to computed $\exp(kA_f\partial_x)U_m^n$. Set $W^{TSM} = W_s + W_f^{\exp}$. Typically $W^{TSM} \approx W$. The error over one unit of time for the split scheme is bounded by

$$\left\| \left((Q_f(k/2)Q_s(k)Q_f(k/2))^{1/k} - \exp(A\partial_x) \right) u \right\|$$
$$\leq \frac{1}{k} \| E_{\text{split}}(k)u + E_s(k)u + 2E_f(k/2)u + 0(k^4) \|$$

For (2.4b), $E_f(k/2) = 0$. From (2.5a),

$$\|E_s(k)u\| \leq \tfrac{1}{6}k(k^2b^3 + h^2b)$$
$$= \tfrac{1}{6}k^3(b^3 + b/\lambda^2).$$

The splitting error is bounded using (2.2),

$$\|E_{\text{split}}(k)u\| \leq \tfrac{1}{6}k^3(a^2b + ab^2)$$
$$\approx \tfrac{1}{6}k^3a^2b,$$

although it may be much smaller for some problems. Since our results depend very much on the size of this error, we will suppose for now that

$$\|E_{\text{split}}(k)u\| \leq \tfrac{1}{6}k^3\sigma$$

for some σ, so that

$$\frac{1}{k}\|E_{\text{split}}(k)u + E_s(k)u\| \leq \tfrac{1}{6}k^2(\sigma + b^3 + b/\lambda^2).$$

In order to obtain accuracy τ we must take

$$k^2 = \frac{6\tau}{\sigma + b^3 + b/\lambda^2}$$

so

$$w(\tau;\lambda) = \lambda W^{TSM}/k^2$$
$$= \lambda(\sigma + b^3 + b/\lambda^2)\frac{W^{TSM}}{6\tau}. \tag{2.10}$$

The optimal stepsize ratio λ now depends on the size of the splitting error and is given by

$$\lambda = \sqrt{\frac{b}{\sigma + b^3}} \tag{2.11}$$

so that

$$w(\tau) = \sqrt{b(\sigma + b^3)}\frac{W^{TSM}}{3\tau} \qquad \text{for the time split method (2.4a,b).}$$

If $\sigma < b^3$ (e.g. when A_f and A_s commute), then (2.11) gives $\lambda \approx$ l/b and

$$W(T) = \frac{b^2 W^{TSM}}{3\tau}. \tag{2.12}$$

When $W^{TSM} \approx W$ this is better than (2.8) by a factor of ϵ^2, meaning greatly improved efficiency. Note that when $\sigma = 0$ the only error incurred is the error in using Lax-Wendroff on the slow scale. From our previous discussion of Lax-Wendroff it is clear why $\lambda = 1/b$ is optimal in this case.

On the other hand, if the splitting error is as bad as (2.0) indicates, then $\sigma = a^2 b$ and $\lambda \approx$ l/a in (2.11) giving

$$w(\tau) = \frac{abW^{TSM}}{3\tau}.$$

This is still an improvement over (2.8), although now only by a factor of ϵ. Note that now λ is chosen appropriate to the fast scale, even though the fast part of the problem is solved exactly, in order to reduce the error due to splitting. Indeed, if we try to use $\lambda = 1/b$ when $\sigma = a^2 b$, we obtain no improvement over (2.8). For this reason it is advisable to always use small time steps with the time-split method (2.4a,b) unless $E_{\text{split}}(k)$ is known to be very small, in which case even greater efficiency is achieved by using larger timesteps.

Now consider the method (2.4a,c) where Lax-Wendroff is used for both operators. In this case $W^{TSM} = W_s + mW_f$ where W_f is the work requried to apply Lax-Wendroff on the fast scale. We are assuming that $W_f << W_s \approx W$. We will take m $\approx \lambda a$ as suggested in (2.6). Using (2.5c) we find that

$$\frac{1}{k}\|E_{\text{split}}(k)u + E_s(k)u + 2E_f(k/2)u\| \le \tfrac{1}{6}k^2(\sigma + b^3 + b/\lambda^2 + a^3/m^2 + a/X^n).$$

We then obtain

$$w(\tau;\lambda) = \lambda(\sigma + b^3 + b/\lambda^2 + 2a/\lambda^2)\frac{W_s + \lambda a W_f}{6\tau}$$
$$\approx \lambda(\sigma + b^3 + 2a/\lambda^2)\frac{W_s + \lambda a W_f}{6\tau} \quad \text{for (2.4a,c).} \tag{2.13}$$

The optimal λ now depends-on the relation between W_f and W_s and is more difficult to solve for. We will discuss three possible choices of λ: λ = l/a, λ = l/b, and $\lambda = 1/\sqrt{ab}$.

When $\lambda = 1/a$, $m = 1$ and we are simply alternating between $LW(A_f, k)$ and $LW(A_s, k)$. In general we would not expect this to be any more efficient than using the unsplit method $LW(A, k)$. Indeed, we find that

$$w(\tau; 1/a) \approx ((\sigma + b^3)/a + 2a^2)\frac{W_s + W_f}{6\tau}$$
$$\approx a^2\frac{W_s + W_f}{3\tau}$$

regardless of the size of σ. This is better than (2.8) only if $W_s + W_f < W$, which is generally not the case for the problems we are considering. (Note that this is the case, however, in the LOD method, where we alternate between solving one-dimensional implicit problems in different space dimensions.)

For λ = l/b increased efficiency is possible if the splitting error is small. From (2.13),

$$w(\tau; 1/b) = (\sigma/b + 2b^2 + 2ab)\frac{W_s + \frac{a}{b}W_f}{6\tau}$$
$$\approx (\sigma/b + 2ab)\frac{W_s + \frac{a}{b}W_f}{6\tau}.$$

Suppose that $W_f + \epsilon W_s = \gamma W$ for some $\gamma \leq 1/2$. Then $W_s + \frac{a}{b} W_f = \frac{\gamma a}{b} W$ and so

$$w(\tau; 1/b) = (a\sigma/b^2 + 2a^2)\frac{\gamma W}{6\tau}. \tag{2.14}$$

This is better than (2.8) whenever $\sigma < ab^2/\gamma$. In this case the time-split method is more efficient. For example, if $\sigma = 0$, (2.14) is better than (2.8) by a factor of 7.

Unfortunately, if $\sigma = a^2 b$ as in (2.9), then

$$w(\tau; 1/b) \approx \frac{a^2(W_s + \frac{a}{b} W_f)}{3\tau}$$

which is no better than (2.8) and may bc worse if $W_f > \epsilon W$.

Now consider an intermediate stepsize ratio, $\lambda = 1/\sqrt{ab}$. From (2.13),

$$\begin{aligned} w(\tau; 1/\sqrt{ab}) &\approx \frac{1}{\sqrt{ab}}(\sigma + 2a^2 b)\frac{W_s + \sqrt{a/b}\, W_f}{6\tau} \\ &\leq a^{3/2} b^{1/2} \frac{W_s + \sqrt{a/b}\, W_f}{2\tau} \\ &= \sqrt{\epsilon} a^2 \frac{W_s + \sqrt{a/b}\, W_f}{2\tau} \end{aligned}$$

regardless of the size of σ. This is better than (2.8) if $W_f + \sqrt{\epsilon} W_s \leq \frac{2}{3} W$, which will generally be true.

We conclude that the method (2.4a,c) is more efficient when λ is chosen correctly. If the splitting error is known to be small then $\lambda = 1/b$ can be used. Otherwise smaller timesteps should be used, e.g. $\lambda = 1/\sqrt{ab}$. Very small timesteps, $\lambda = $ l/a, should never be used.

Here we have not dealt with the advantages of the split scheme resulting from the possibility of choosing the stepsize ratio on each scale so that the k^3 and kh^2 terms in each of the Lax-Wendroff errors nearly cancel out. When this can be done, the splitting may be even more advantageous than indicated here.

Block triangular systems. Since the efficiency of the split scheme is limited primarily by the splitting error, it is interesting to investigate how this error depends on the coupling between fast and slow scales in a simple model system. Suppose that the matrix A is of the form (1.10) with $\|A_{12}\| \approx \alpha \leq 1$ and that the splitting (1.11) is used. Here A_{12} is the coupling between fast and slow scales. If $A_{12} = 0$, the problem is uncoupled and $E_{\text{split}}(k) = 0$. In general, from (2.2),

$$E_{\text{split}}(k) = -\frac{k^3}{6}\begin{bmatrix} 0 & \frac{1}{4\epsilon} A_{11}(\frac{1}{\epsilon} A_{11} A_{12} - 2A_{12}A_{22}) \\ 0 & I \end{bmatrix} \partial_x^3 + O(k^4).$$

Thus $\|E_{\text{split}}(k)u\| \approx \alpha k^3/24\epsilon^2$. The' efficiency- of the splitting depends on the size of α. In the notation used above, we have

$$a = 1/\epsilon, \qquad b = 1, \qquad \sigma = \tfrac{1}{4}\alpha a^2 b.$$

For unsplit Lax-Wendroff, (2.8) gives

$$w(\tau) = \frac{1}{\epsilon^2}\frac{W}{3\tau}. \tag{2.15}$$

The time-split method (2.4a,b) is always more efficient if we choose

$$\lambda \approx (1 + \tfrac{1}{4}\alpha a^2 b)^{-1/2}.$$

For example, if $\alpha \approx 1$ we should use $\lambda \approx 2/a = 2\epsilon$ in order to reduce (2.15) by a factor of ϵ. The maximum efficiency indicated in (2.12) is achievable only if $\alpha \leq \epsilon^2$, in which case taking $\lambda = 1$ reduces (2.15) by a factor or ϵ^2.

Reducing the splitting error. For block triangular systems in which A_{12} is not sufficiently small, it is possible to reduce the coupling through a change of variables so that the optimal efficiency can be achieved. A change of variables amounts to replacing u by $\bar{u} = Bu$ for some nonsingular matrix B. The system $u_t = Au$, then becomes $\bar{u}_t = BAB^{-1}\bar{u}_x$. Clearly, if B is chosen to be the eigenvector matrix of A then the problem completely decouples into independent scalar equations. We are seeking something less expensive which only decouples the fast and slow scales. Thus we want a matrix B such that

$$BAB^{-1} = \begin{bmatrix} \frac{1}{\epsilon}C_{11} & 0 \\ 0 & C_{22} \end{bmatrix} \tag{2.16}$$

with $\|C_{11}\| \approx \|C_{22}\| \approx 1$. In the block triangular case, it suffices to consider B of the form

$$B = \begin{bmatrix} I & B_{12} \\ 0 & I \end{bmatrix}, \qquad B^{-1} = \begin{bmatrix} I & -B_{12} \\ 0 & I \end{bmatrix}.$$

Then

$$BAB^{-1} = \begin{bmatrix} \frac{1}{\epsilon}A_{11} & -\frac{1}{\epsilon}A_{11}B_{12} + A_{12} + B_{12}A_{22} \\ 0 & A_{22} \end{bmatrix}$$

and so B_{12} should be chosen to solve

$$\frac{1}{\epsilon}A_{11}B_{12} - B_{12}A_{22} = A_{12} \tag{2.17}$$

in order to completely decouple the fast and slow scales.

In the present context solving for B_{12} from (2.17) is not worthwhile. In order to achieve optimal efficiency we need only reduce the coupling by one or two factors of ϵ. Further reductions do not gain anything once the Lax-Wendroff errors dominate. This suggests taking

$$B_{12} = \epsilon A_{11}^{-1}A_{12} \tag{2.18}$$

so that

$$BAB^{-1} = \begin{bmatrix} \frac{1}{\epsilon}A_{11} & A_{12}^{(1)} \\ 0 & A_{22} \end{bmatrix}$$

where

$$A_{12}^{(1)} = \epsilon A_{11}^{-1}A_{12}A_{22}.$$

We now have $\|A_{12}^{(1)}\| \approx \epsilon\alpha$ provided $\|A_{11}^{-1}\| \approx 1$. The coupling is thus reduced by a factor of ϵ through the use of a very simple change of variables. The above process can be repeated to obtain additional factors of ϵ. This change of variables has been suggested by Kreiss[9] in a similar context.

For full systems of the form

$$A = \begin{bmatrix} \frac{1}{\epsilon}A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

we can obtain a similar reduction in the size of both off-diagonal blocks and again reduce the splitting error by several orders of magnitude. In this case we consider B of the form

$$B = \begin{bmatrix} I & K \\ 0 & I \end{bmatrix}\begin{bmatrix} I & 0 \\ L & I \end{bmatrix} = \begin{bmatrix} I + KL & K \\ L & I \end{bmatrix}.$$

It is easy to verify that the lower corner of A is annihilated by taking L to satisfy

$$\frac{1}{\epsilon}LA_{11} - A_{22}L - LA_{12}L + A_{21} = 0.$$

The matrix K can then be chosen as before to remove the remaining upper corner. This results in a system of the form (2.16). This particular transformation is discussed more completely by O'Malley and Anderson[14]. Again, however, we are not interested here in completely annihilating the corners, but rather in reducing them by a factor of ϵ. This is easily accomplished by taking

$$K = \epsilon A_{11}^{-1}A_{12}$$
$$L = -\epsilon A_{21}A_{11}^{-1}.$$

Example 8.1 in section 8 illustrates the use of the change of variables for a triangular system.

3. Stability of the time-split method

In this section we investigate the stability of the time-split method when applied to a constant coefficient problem on the entire real line, $-\infty < x < \infty$ or, alternatively, on a finite interval with periodic boundary conditions. When $Q_f^2(k/2) = Q_f(k)$, as is true for the splittings (2.4), for example, Cauchy stability of the Strang splitting (1.8) is equivalent to stability of the first order splitting

$$U^{n+1} = Q_s(k)Q_f(k)U^n. \tag{3.1}$$

For simplicity we restrict our attention to this splitting, and set Q(k) $= Q_s(k)Q_f(k)$.

In general the stability of $Q_s(k)$ and $Q_f(k)$ does not imply stability of $Q(k)$. Instead stability must bc checked directly. In fact, (3.1) can be unstable even when $Q_f(k)$ and $Q_s(k)$ are exact solution operators for well-posed hyperbolic problems as the following example shows.

Example 3.1. Let

$$A_f = \begin{bmatrix} 1 & \mu \\ 0 & -1 \end{bmatrix}, \qquad A_s = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Then the problems $u_t = A_f u_x$, $u_t = A_s u_x$ and $u_t = (A_f + A_s)u_x$ are all well-posed, strictly hyperbolic problems for any value of the parameter μ. Let

$$Q_f(k) = \exp(kA_f\partial_x), \qquad Q_s(k) = \exp(kA_s\partial_x).$$

and let $G_f(\xi, k/2)$ and $G_s(\xi, k)$ be the corresponding amplification matrices. For the exact solution operators,

$$G_f(\xi, k/2) = \exp(ik\xi A_f)$$
$$= \begin{bmatrix} e^{ik\xi} & \text{pi sin } k\xi \\ 0 & e^{-ik\xi} \end{bmatrix}$$

and

$$G_s(\xi, k) = \exp(ik\xi A_s)$$
$$= \begin{bmatrix} \cos k\xi & i \sin k\xi \\ i \sin k\xi & \cos k\xi \end{bmatrix}.$$

We have $\rho(G_f(\xi, k/2)) = \rho(G_s(\xi, k)) = 1$ for all ξ and k. On the other hand, the amplification matrix $G(\xi, k)$ for the time-split method (3.1) has $\rho(G(\xi, k)) = 1$ for all ξ and k only if $|\mu| \leq 2$. When $|\mu| > 2$, the method is unstable. Figure 3.1 shows graphs of $\rho(G(\xi, k))$ for $\mu = 5,10$.

In spite of this example, there are some very important classes of splittings for which the individual stability of $Q_f(k)$ and $Q_s(k)$ does imply the stability of $Q(k)$. It is useful to delineate such classes, since the stability of $Q_f(k)$ and $Q_s(k)$ is often easy to determine, whereas the stability of Q(k) may be quite tedious to determine directly.

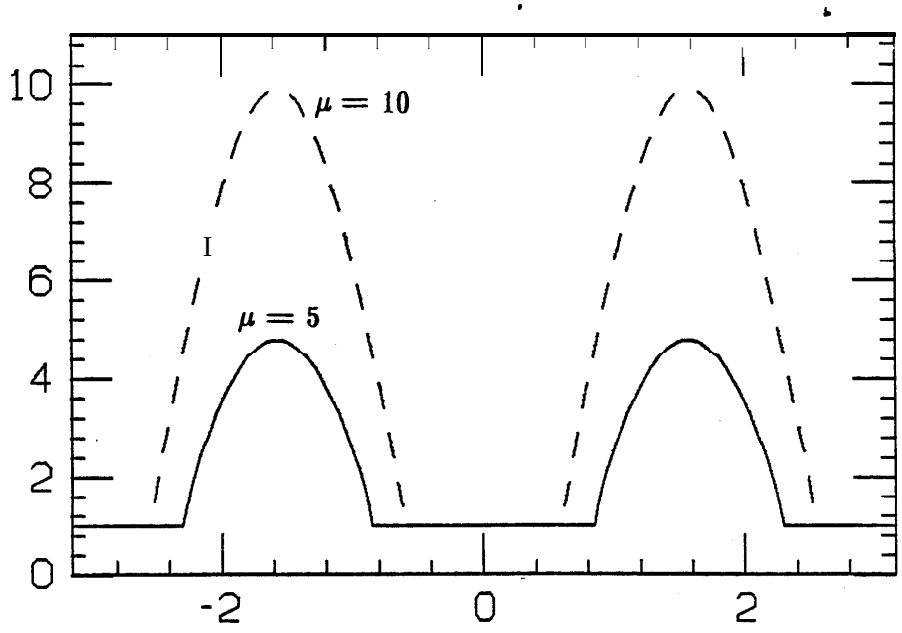

Figure 3.1. Spectra1 radius of the amplification matrix $G(\xi, k)$ *of* example *3.1* for $\mu = 5, 10$, *as a function of* ξk *between* $-\pi$ *and* π.

Block triangular systems. One such case is the block triangular system of equations

$$\begin{bmatrix} u \\ v \end{bmatrix}_t = \begin{bmatrix} \frac{1}{\epsilon} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}_x.$$

with the splitting (1.11). The solution v does not depend on u. In solving for u, the computed v enters essentially as a forcing function. The schemes $Q_s(k)$ and $Q_f(k)$ will be of the form

$$Q_s(k) = \begin{bmatrix} I & Q_{12}(k) \\ 0 & Q_{22}(k) \end{bmatrix}, \qquad Q_f(k) = \begin{bmatrix} Q_{11}(k) & 0 \\ 0 & I \end{bmatrix}. \tag{3.2}$$

Suppose that $Q_{11}(k)$ and $Q_{22}(k)$ are stable schemes, and in particular, that there exist a norm $\|\cdot\|$ and a constant $\alpha \geq 0$ such that

$$\|Q_{11}(k)\| < 1 + \alpha k \qquad \text{for all } k \text{ sufficiently small.} \tag{3.3}$$

All of the following estimates will be in this norm. We also suppose that

$$\|Q_{12}(k)V\| \leq kM\|D_+V\| \tag{3.4}$$

for some constant M. These assumptions are satisfied for the methods (2.4) *provided the Lax-Wendroff operators are stable. WC then have the following theorem.

THEOREM 3.1. *Suppose $Q_f(k)$ and $Q_s(k)$ are stable schemes as above. Then the split scheme $Q_s(k)Q_f(k)$ is stable for smooth initial data V^0. More precisely, we obtain bounds for the solution which depend on a discrete Sobolev norm of the initial data,*

$$\|U^n\| \le K_T(\|U^0\| + \|D_+V^0\|) \tag{3.5a}$$
$$\|V^n\| \le \tilde{K}_T\|V^0\| \tag{3.5b}$$

for $nk \le T$. *Here K_T and $\tilde{K}_T$ are constants depending only on the fixed time T.*

Proof. When the full scheme $\vec{U}^{n+1} = Q_s(k)Q_f(k)\vec{U}^n$ is written out we obtain

$$U^{n+1} = Q_{11}(k)U^n + Q_{11}(k)Q_{12}(k)V^n \tag{3.6a}$$
$$V^{n+1} = Q_{22}(k)V^n \tag{3.6b}$$

The bound (3.5b) follows immediately from (3.6b) and the stability of $Q_{22}(k)$. Moreover, by linearity, an identical bound holds for the linear combination of solutions D_+V^n, i.e.,

$$\|D_+V^n\| \le \tilde{K}_T\|D_+V^0\|.$$

Using this together with (3.4) in (3.6a) gives

$$\|U^{n+1}\| \le \|Q_{11}(k)\|(\|U^n\| + kM\tilde{K}_T\|D_+V^0\|).$$

When iterated n times this gives

$$\|U^n\| \le \|Q_{11}(k)\|^n\|U^0\| + kM\tilde{K}_T(\|Q_{11}(k)\|^{n-1} + \|Q_{11}(k)\|^{n-2} + \cdots + \|Q_{11}(k)\| + 1)\|D_+V^0\|. \tag{3.7}$$

BY (3.3), $\|Q_{11}(k)\|^n \le (1 + \alpha k)^n \le e^{\alpha T}$ if $nk \le T$. Using this in (3.7) gives

$$\|U^n\| \le e^{\alpha T}(\|U^0\| + TM\tilde{K}_T\|D_+V^0\|)$$

for $nk \le T$, which is of the desired form (3.5a). ∎

Simultaneously normalizable matrices. Stability also follows dircc tly when A_f and A_s . arc normal matrices (a normal matrix is one which commutes with its transpose). This includes, for example, symmetric matrices and scalar problems. In fact, it suffices that A_f and A, be simultaneously normalizable, i.e., that there exist some nonsingular matrix S such that SA_sS^{-1} and $SA_f S^{-1}$ are both normal. Thus, the case of simultaneously diagonalizable A_f and A, is also covered. This is a consequence of the following, even more general, theorem.

THEOREM 3.2. Let $A1, A_2, \ldots, A_m$, be constant matrices. Approximate each solution operator $\exp(k_jA_j\partial_x)$ by some operator $Q_j(k_j)$ with amplification matrix $G_j(\xi)$. Suppose there exists some norm $\|\cdot\|$ for which

$$\|G_j(\xi)\| \le 1 \qquad \forall\xi, \quad j = 1, 2, \ldots, m. \tag{3.3}$$

Then the scheme

$$U^{n+1} = Q_1(k_1)Q_2(k_2)\cdots Q_m(k_m)U^n \tag{3.4}$$

is stable.

Proof. Let $G(\xi) = G_1(\xi)G_2(\xi)\cdots G_m(\xi)$. Then powers of $G(\xi)$ are uniformly bounded in the norm $\|\cdot\|$ since

$$\|G^n(\xi)\| \le \|G(\xi)\|^n \le (\|G_1(\xi)\| \cdots \|G_m(\xi)\|)^n \le 1.$$

It follows. that (3.4) is stable. ∎

COROLLARY *Suppose there exists some nonsingular matrix* S *such that* SA_jS^{-1} is *normal for* $j = 1, 2, \ldots, m$ *and that the amplification matrices* $G_j(\xi)$ *satisfy*

$$\text{i) } SG_j(\xi)S^{-1} \text{ is also normal for all } \xi, \quad j = 1,2,\ldots,m \qquad \text{ii) } \|SG_j(\xi)S^{-1}\| \le 1 \quad \forall \xi, \quad j = 1,2,\ldots,m \tag{3.5}$$

Then the scheme (3.4) is stable.

Remark: Condition $(3.5ii)$ is satisfied if $Q_j(k_j)$ is the exact solution operator or one or more steps of Lax-Wendroff.

Proof. Since the 2-norm of a normal matrix is equal to its spectral radius, conditions (3.5) give

$$\|SG_j(\xi)S^{-1}\|_2 = \rho(SG_j(\xi)S^{-1}) = \rho(G_j(\xi)) \le 1.$$

It follows that the hypothesis of Theorem 3.2 is satisfied in the norm $\|\cdot\|$ defined by

$$\|A\| = \|SAS^{-1}\|_2.$$

This completes the proof. ∎

4. The Leapfrog Duhamel method.

As mentioned in the introduction, the time-split method does not immediately lend itself to use with multi-level difference schemes. We now present a new method with the same basic philosophy as the time-split method but which uses leapfrog on the slow time scale.

Using Duhamel's principle (i.e., variation of parameters) we can write the solution to (1.3) as

$$u(x, t + k) = \exp(2kA_f\partial_x)u(x, t - k) + \int_{t-k}^{t+k} \exp((t + k - \tau)A_f\partial_x)A_s u_x(x, \tau)\, d\tau.$$

If we now approximate the integral by the midpoint rule we obtain

$$\begin{aligned} u(x, t + k) &\approx \exp(2kA_f\partial_x)u(x, t - k) + 2k \exp(kA_f\partial_x)A_s u_x(x, t) \\ &= \exp(kA_f\partial_x)[\exp(kA_f\partial_x)u(x, t - k) + 2kA_s u_x(x, t)]. \end{aligned}$$

Replacing $u_x(x, t)$ by the standard centered difference operator and approximating $\exp(kA_f\partial_x)$ by $Q_f(k)$ gives the ***Leapfrog Duhamel method,***

$$U_m^{n+1} = Q_f(k)[Q_f(k)U_m^{n-1} + \frac{k}{h}A_s(U_{m+1}^n - U_{m-1}^n)]. \tag{4.1}$$

The term inside the brackets is essentially leapfrog for the problem $u_t = A_s u_x$ since $Q_f(k)U^{n-1} \approx \exp(-kA_s\partial_x)U^n$. If $Q_f(k)$ is an $O(k^3)$ approximation to $\exp(kA_f\partial_x)$ then (4.1) provides an $O(k^3)$ accurate approximate solution, even for noncommuting A_f and A_s. This will be shown in section 5 where the method is analyzed in more detail.

We pay a price for using a scheme involving three time levels, since (4.1) requires two applications of the operator $Q_f(k)$. One of these is needed only to provide the proper values at time $n - 1$. Nevertheless, this method may be useful, particularly in cases where $\exp(kA_f\partial_x)$ is known exactly and thus is easy to apply.

5. Accuracy of the Leapfrog Duhamel method

The Leapfrog Duhamel scheme can be analyzed in terms of the error in the midpoint rule directly from its derivation. We prefer to build upon the results in section 2 by rewriting Leapfrog Duhamel as a splitting.

First consider the standard leapfrog scheme on $u_t = A_s u_x$,

$$U^{n+1} = U^{n-1} + 2kA_sD_0U^n.$$

The truncation error is given by

$$\begin{aligned}&[u(x, t-k) + 2kA_sD_0u(x, t)] - u(x, t+k)\\ &\quad = [(I + 2kA_sD_0\exp(kA_s\partial_x)) - \exp(2kA_s\partial_x)]u(x, t-k).\end{aligned}$$

For conformity with section 2, we define the operator $Q_s(2k)$ by

$$Q_s(2k) = I + 2kA_sD_0\exp(kA_s\partial_x).$$

Note that this is not the actual finite difference operator for leapfrog, since in general U^n is not exactly equal to $\exp(kA_s\partial_x)U^{n-1}$, but it is the proper operator for computing the local truncation error, in which U^{n-1} and U^n are replaced by the true solution values. We can now define the truncation error operator for leapfrog on stepsize k by

$$E_s^{LF}(k) = Q_s(2k) - \exp(2kA_s\partial_x) = O(k^3). \tag{5.1}$$

The Leapfrog Duhamel scheme is

$$U^{n+1} = Q_f(k)\big(Q_f(k)U^{n-1} + 2kA_sD_0U^n\big) \tag{5.2}$$

where $Q_f(k)$ is some approximation to $\exp(kA_f\partial_x)$ with error operator

$$E_f(k) = Q_f(k) - \exp(kA_f\partial_x) = O(k^3).$$

To obtain the truncation error for Leapfrog Duhamel we replace U^{n-1} and U^n by $u(x, t-k)$ and $u(x, t)$ in (5.2). The right hand side then becomes

$$\begin{aligned}&Q_f(k)[I + 2kA_sD_0\exp(kA\partial_x)Q_f^{-1}(k)]Q_f(k)u(x,t-k)\\ &\quad = Q_f(k)[I + 2kA_sD_0\exp(kA_s\partial_x) + 2kA_sD_0(\exp(kA\partial_x)Q_f^{-1}(k)\\ &\qquad -\exp(kA_s\partial_x))]Q_f(k)u(x,t-k) \qquad (5.3)\\ &\quad = [Q_f(k)Q_s(2k)Q_f(k) + 2kQ_f(k)A_sD_0(\exp(kA\partial_x)\\ &\qquad -\exp(kA_s\partial_x)\exp(kA_f\partial_x) - \exp(kA\partial_x)E_f(k))]u(x,t-k).\end{aligned}$$

Thus the Leapfrog Duhamel operator can be viewed as a splitting of the form $Q_f(k)Q_s(2k)Q_f(k)$ plus some additional error terms which are $O(k^3)$. Let $E^I_{\text{split}}(k)$ denote the error operator for the first order accurate splitting

$$E^I_{\text{split}}(k) = \exp(kA\partial_x) - \exp(kA_s\partial_x)\exp(kA_f\partial_x) = O(k^2).$$

Observing that

$$\exp(kA_s\partial_x)E_f(k) = O(k^3),$$
$$Q_f(k) = I + O(k),$$
$$D_0 = \partial_x + O(k^2),$$

the operator in (5.3) becomes

$$Q_f(k)Q_s(2k)Q_f(k) + 2kA_s\partial_x E^I_{\text{split}}(k) + O(k^4).$$

Using (2.3) we obtain an expression for the truncation error operator for Leapfrog Duhamel,

$$\begin{aligned} E^{LFD}(k) &= (Q_f(k)Q_s(2k)Q_f(k) + 2kA_s\partial_x E^I_{\text{split}}(k) + O(k^4)) \\ &\quad - \exp(2kA\partial_x) \\ &= E_{\text{split}}(2k) + E^{LF}_s(k) + 2E_f(k) + 2kA_s\partial_x E^I_{\text{split}}(k) + O(k^4). \end{aligned}$$

For A_s and A_f constant we have

$$E^I_{\text{split}}(k) = \frac{1}{2}k^2(A_fA_s - A_sA_f)\partial_x^2 + O(k^3)$$

so

$$\begin{aligned} &E_{\text{split}}(2k) + 2kA_s\partial_x E^I_{\text{split}}(k) \\ &\quad = -\tfrac{1}{3}k^3(A_f^2A_s - 2A_fA_sA_f + A_sA_f^2 \\ &\qquad + A_s^2A_f + A_sA_fA_s - 2A_fA_s^2)\partial_x^3. \end{aligned}$$

The splitting error in Leapfrog Duhamel is thus roughly 8 times as large as the corresponding error in the time split method with Lax-Wendroff. The work comparisons of section 2 can be repeated for Leapfrog Duhamel with similar results.

6. Stability of the Leapfrog Duhamel method

At present the stability analysis for Leapfrog Duhamel covers only the case in which A_f and A_s are simultaneously diagonalizable,

$$XA_fX^{-1} = M_f, \qquad XA_sX^{-1} = M_s$$

where M_f and M_s are diagonalizable matrices. We assume that $Q_f(k)$ is stable and is also diagonalized by X. This is true for $Q_f(k) = \exp(kA_f\partial_x)$ or for $Q_f(k) = (LW(A_f, k/m))^m$ with $\rho(A_f)k/mh \leq 1$. Let $q_f(k)$ be a single diagonal element of $XQ_f(k)X^{-1}$ and μ_s a diagonal element of M_s. It suffices to consider the scalar equation

$$U^{n+1} = q_f^2(k)U^{n-1} + 2kq_f(k)\mu_s D_0 U^n. \tag{6.1}$$

Let $g_f(\xi)$ be the amplification factor corresponding to $q_f(k)$. By assumption, $|g_f(\xi)| \leq 1$ for all ξ.

THEOREM 6.1. *Suppose $|\lambda\mu_s| \leq 1$, where $\lambda = k/h$. Then the amplification factor $g(\xi)$ for the scheme (6.1) satisfies*

$$|g(\xi)| = |g_f(\xi)|.$$

Proof. The amplification factor is dcrivctl by letting

$$U_m^n = g^n(\xi)e^{i\xi mh}$$

in (6.1). We obtain the equation

$$s(E) = g_f^2(\xi)g^{-1}(\xi) + 2i\lambda g_f(\xi)\mu_s \sin \xi h$$

which can be rewritten as

$$(g(\xi)g_f^{-1}(\xi))^2 - 2i\lambda\mu_s \sin \xi h(g(\xi)g_f^{-1}(\xi)) - 1 = 0.$$

Solving this quadratic equation yields

$$g(\xi)g_f^{-1}(\xi) = i\lambda\mu_s \sin \xi h \pm \sqrt{1 - \lambda^2\mu_s^2 \sin^2 \xi h}.$$

If $|\lambda\mu_s| \leq 1$, the square root is real and so

$$|g(\xi)g_f^{-1}(\xi)|^2 = 1$$

and hence

$$|g(\xi)| = |g_f(\xi)|$$

as claimed. ∎

Note that when the exact solution operator is used for $q_f(k)$ we have $|g_f(\xi)| = 1$ and hence $|g(\xi)| = 1$ for all ξ. In this case Leapfrog Duhamel is nondissipative.

7. Boundary data for the intermediate solutions.

For general initial boundary value problems we must be able to generate the appropriate boundary values for the intermediate solutions which arise in the use of a split scheme. We have developed a general methodology for defining the proper boundary data which will be illustrated here for constant coefficient problems at an inflow boundary. More general problems can also be handled, as will be reported on elsewhere. The procedure will be demonstrated for the time-split method (2.4a,b), but can also be used- for the other methods previously described.

First consider the scalar problem

$$\begin{aligned} u_t &= -(1+\epsilon)u_x \qquad x \geq 0, t \geq 0 \\ u(0,t) &= g(t) \qquad\quad t \geq 0 \end{aligned} \tag{7.1}$$

with the splitting

$$A_1 = -1, \quad A_2 = -\epsilon.$$

Take $k = 2h$ and use the method of characteristics solution for A_f and Lax-Wendroff on A_s. There is no need to use a Strang-type splitting, since the operators commute, and thus the split scheme is simply

$$\begin{aligned} U_m^* &= U_{m-2}^n \qquad m = 2,3,\ldots \\ U_m^{n+1} &= U_m^* - \epsilon(U_{m+1}^* - U_{m-1}^*) + 2\epsilon^2(U_{m+1}^* - 2U_m^* + U_{m-1}^*) \\ &\qquad m = 1,2,\ldots \end{aligned} \tag{7.2}$$

The value of U_0^{n+1} is given by the boundary conditions,

$$U_0^{n+1} = g(t_{n+1}).$$

For the splitting (7.2) we must also provide U_0^* and U_1^*. In general with $k = ph$ for some integer $p \geq 1$, we would need to supply $U_0^*, U_1^*, \ldots, U_{p-1}^*$.

In order to generate boundary data we consider U_m^* as an approximation to $u^*(x_m, t_{n+1})$ where the continuous function $u^*(x, t)$ satisfies

$$\begin{aligned} u_t^* &= -u_x^* \qquad x \geq 0,\ t \geq t_n \\ u^*(x, t_n) &= u(x, t_n) \qquad x \geq 0. \end{aligned} \tag{7.3}$$

Then, using the differential equations governing u and u^*, we can express U_0^* and U_1^* in terms of g(t). Consider U_0^*. We want

$$\begin{aligned} U_0^* &= u^*(0, t_n + k) \\ &= u^*(0, t_n) + k u_t^*(0, t_n) + \tfrac{1}{2}k^2 u_{tt}^*(0, t_n) + \ldots \\ &= u^*(0, t_n) - k u_x^*(0, t_n) + \tfrac{1}{2}k^2 u_{xx}^*(0, t_n) + \ldots \end{aligned} \tag{7.4}$$

Mere we used (7.3) to express u_t^* in terms of u_x^*. But since $u^*(x, t_n) = u(x, t_n)$ for all x, this relation can be differentiated with respect to x, giving $u_x^*(x, t_n) = u_x(x, t_n)$ and similarly for higher derivatives. So (7.4) becomes

$$U_0^* = u(0, t_n) - k u_x(0, t_n) + \tfrac{1}{2}k^2 u_{xx}(0, t_n) + \ldots.$$

We can now use the original equations (7.1) governing u to rewrite this in terms of t-derivatives of u. Since

$$\partial_x^j u = \left(\frac{-1}{1+\epsilon}\right)^j \partial_t^j u \qquad j \geq 0$$

we obtain

$$\begin{aligned} U_0^* &= u(0, t_n) + \tfrac{k}{1+\epsilon} u_t(0, t_n) + \tfrac{1}{2}\left(\tfrac{k}{1+\epsilon}\right)^2 u_{tt}(0, t_n) \cdots \\ &= u(0, t_n - k/(1 \ldots \\ &= g(t_n - k/(1 - \epsilon)). \end{aligned} \tag{7.5}$$

This is the desired boundary data.

For such a simple example it is easy to verify that this is the correct boundary value. According to the scheme (7.2) we would really like

$$U_0^* = U_{-2}^n = u(-ih, t_n).$$

Of course u is not officially defined for x < 0, but using the differential equation (7.1) it can easily be extended from the boundary. Since (7.1) has characteristics with slope $1/(1 + E)$, we Gnd that

$$u(-2h, t_n) = u(0, t_n + 2h/(1+\epsilon)) = g(t_n + k/(1+\epsilon))$$

exactly as in (7.5) .

We can compute U_1^* in the same manner. We want

$$\begin{aligned} U_1^* &= u^*(h, t_{n+1}) \\ &= u^*(0, t_{n+1/2}) \qquad \text{where } t_{n+1/2} = t_n + k/2. \end{aligned}$$

We now proceed as before,

$$\begin{aligned} U_1^* &= u^*(0,t_n) - \tfrac{1}{2}ku_t^*(0,t_n) + \tfrac{1}{8}k^2u_{tt}^*(0,t_n) - \cdots \\ &= u(0,\cdots) - \tfrac{1}{2}ku_x(0,t_n) + \tfrac{1}{8}k^2u_{xx}(0,t_n) - \cdots \\ &= u(0,t_n) - \tfrac{1}{2}\tfrac{k}{1+\epsilon}u_t(0,t_n) + \tfrac{1}{8}\left(\tfrac{k}{1+\epsilon}\right)u_{tt}(0,t_n) - \cdots \\ &= g(t_n - \tfrac{1}{2}k/(1-\epsilon)). \end{aligned} \tag{7.6}$$

To summarize our procedure, we switched from t-derivatives of u^* to x-derivatives of u^*. Since these were evaluated at time t_n, they were identical to the corresponding x-derivatives of u. We then switched back to t-derivatives of u along the boundary, which allowed us to use the known boundary conditions for u. Clearly this procedure will not work so neatly when we deal with variable coefficients, systems of equations, or inflow-outflow boundaries. Nonetheless, these same ideas, combined with a little ingenuity, lead to sufficiently accurate approximate boundary conditions for a wide variety of problems.

Constant coefficient systems. Next consider the system of equations

$$\begin{aligned} u_t &= Au_x \equiv (A_f + A_s)u_x \qquad x \geq 0, \quad t \geq 0 \\ u(0,t) &= g(t) \qquad\qquad t \geq 0. \end{aligned} \tag{7.7}$$

We assume that A and A_f have strictly negative eigenvalues. In general A_f and A, do not commute, so we will have to use a Strang-type splitting. There will be at least two intermediate solutions, say

$$\begin{aligned} U^* &\approx \exp(\tfrac{1}{2}kA_f\partial_x)U^n \\ U^{**} &\approx \exp(kA_s\partial_x)\exp(\tfrac{1}{2}kA_f\partial_x)U^n. \end{aligned} \tag{7.8}$$

Of course there may be many more if $\exp(\frac{1}{2}kA_f\partial_x)$ is itself approximated by several steps of Lax-Wendroff, but they can be handled similarly. The general principle should be clear from considering (7.8).

Again let $u^*(x,t)$ be a continuous function satisfying

$$\begin{aligned} u_t^* &= A_f u_x^* \qquad x \geq 0, \quad t \geq t_n \\ u^*(x,t_n) &= u(x,t_n) \quad x \geq 0. \end{aligned} \tag{7.9}$$

We then want

$$\begin{aligned} U_0^* &= u^*(0,t_{n+1/2}) \\ &= u^*(0,t_n) + \tfrac{1}{2}ku_t^*(0,t_n) + \tfrac{1}{8}k^2u_{tt}^*(0,t_n) + \cdots \\ &= u(0,t_n) + \tfrac{1}{2}kA_fu_x(0,t_n) + \tfrac{1}{8}k^2A_f^2u_{xx}(0,t_n) + \cdots \\ &= u(0,t_n) + \tfrac{1}{2}kA_fA^{-1}u_t(0,t_n) + \tfrac{1}{8}k^2A_f^2A^{-2}u_{tt}(0,t_n) + \cdots \\ &= g(t_n) + \tfrac{1}{2}kA_fA^{-1}g'(t_n) + \tfrac{1}{8}k^2A_f^2A^{-2}g''(t_n) + \cdots. \end{aligned} \tag{7.10}$$

We assume that the boundary is non-characteristic so that A is invertible. In general U_0^* must now be approximated by the first few terms of (7.10). If we keep only the first two terms we will have boundary data with $O(k^2)$ errors. This is sufficient to retain the $O(k^2)$ global accuracy of Lax-Wendroff (see Gustafsson[7]). It may, however, increase the error constant considerably and partly offset the benefit obtained by using the split scheme. Consider, for example, a case in which

$\|A_f\| \approx 1$ and $\|A_s\| \approx \epsilon$. In this case the error $E^{TSM}(k)u$ is like ϵk^3 at most and the resulting global error, assuming u is smooth, will be like ϵk^2. In order to achieve the same accuracy in the boundary data we will have to include the third terrn of (7.10) as well (or at least its dominant part). In some such cases it happens that

$$A_f^j A^{-j} = \text{I} + \text{O(c)} \quad \text{for j} = 1, 2, \ldots .$$

We can then retain $O(\epsilon k^2)$ accuracy simply by taking

$$U_0^* = g(t_{n+1/2}) + \tfrac{1}{2}k(A_f A^{-1} - I)g'(t_n).$$

This will be illustrated in Example 7.1 .

Now to find boundary values fo U^{**}. The easiest way to proceed is to note that

$$U^{**} = \exp(-\tfrac{1}{2}kA_f\partial_x)U^{n+1}$$

which prornp ts us to define $u^{**}(x, t)$ as the continuous solution to

$$\begin{aligned} u_t^{**}(x, t) &= A_f u_x^{**}(x, t) \qquad x \geq 0, \quad t \leq t_{n+1} \\ u^{**}(x, t_{n+1}) &= u(x, t_{n+1}) \qquad x \geq 0. \end{aligned} \tag{7.11}$$

We now solve this backwards in time for

$$U_0^{**} = u^{**}(0, t_{n+1/2}).$$

Proceeding as in the derivation of (7.10) we obtain

$$\begin{aligned} U_0^{**} &= g(t_{n+1}) - \tfrac{1}{2}kA_f A^{-1}g'(t_{n+1}) + \tfrac{1}{8}k^2 A_f^2 A^{-2}g''(t_{n+1}) + \ldots \\ &\approx g(t_{n+1/2}) - \tfrac{1}{2}k(A_f A^{-1} - I)g'(t_{n+1}). \end{aligned}$$

Example 7.1 Consider

$$\begin{aligned} \begin{bmatrix} \text{U} \\ \text{V} \end{bmatrix}_t &= \begin{bmatrix} -1 & \epsilon_1 \\ \epsilon_2 & -2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}_x \qquad 0 \leq x \leq 1, \quad t \geq 0 \\ \vec{u}(x, 0) &= f(x) \qquad 0 \leq x \leq 1 \\ \vec{u}(0, t) &= g(t) \qquad t \geq 0 \end{aligned}$$

where 3 = (u, $v)^T$. We have. chosen a strip problem to illustrate that outflow boundaries are frequently trivial to handle with a split method. Take

$$A_f = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}, \qquad A_s = \begin{bmatrix} 0 & \epsilon_1 \\ \epsilon_2 & 0 \end{bmatrix}$$

For this problem the splitting error is

$$E_{\text{split}}(k) = -\tfrac{1}{6}k^3 \begin{bmatrix} -\epsilon_1\epsilon_2 & \tfrac{1}{4}\epsilon_1 \\ \tfrac{1}{4}\epsilon_2 & \epsilon_1\epsilon_2 \end{bmatrix} \partial_x^3.$$

If we use the time-split method (2.4a,b) then, according to (2.11), the optimal stepsize ratio is

$$\lambda \approx \sqrt{\frac{\epsilon}{\frac{1}{4}\epsilon + \epsilon^3}} \approx 2$$

where $\epsilon = \max |\epsilon_j|$. For $k = 2h$ and $h = 1/M$, (2.4a,b) becomes:

$$U_m^* = U_{m-1}^n, \quad m = 1, 2, \ldots, M$$
$$V_m^* = V_{m-2}^n, \quad m = 2, 3, \ldots, M$$

$$\vec{U}_m^{**} = LW(A_s, k)\vec{U}_m^*, \quad m = 1, 2, \ldots, M - 1$$

$$\vec{U}_0^{n+1} = g(t_{n+1})$$
$$U_m^{n+1} = U_{m-1}^{**}, \; m = 1, 2, \ldots, M$$
$$V_m^{n+1} = V_{m-2}^{**}, \; m = 2, 3, \ldots, M$$

Notice that no boundary conditions whatsoever need to be specified at the outflow boundary x = 1. On the inflow side we still need to specify $\vec{U}_0^*$, V_1^*, $\vec{U}_0^{**}$, and V_1^{n+1}. For ths problem,

$$A_f^2 A^{-2} = \frac{1}{(2 - \epsilon_1\epsilon_2)^2}\begin{bmatrix} 4 - 12\epsilon_2\epsilon_2 & 4 + 34\epsilon_1\epsilon_2 \\ \cdots & \cdots \end{bmatrix}$$
$$= I + O(\epsilon).$$

and we can retain $O(\epsilon k^2)$ accuracy by taking

$$\vec{U}_0^* = g(t_{n+1/2}) + \tfrac{1}{2}k(A_f A^{-1} - I)g'(t_n)$$
$$= g(t_{n+1/2}) + \frac{k}{2(2 - \epsilon_1\epsilon_2)}\begin{bmatrix} \epsilon_1\epsilon_2 & \epsilon_1 \\ 2\epsilon_2 & \epsilon_1\epsilon_2 \end{bmatrix} g'(t_n) \tag{7.14}$$

Similarly we use

$$\vec{U}_0^{**} = g(t_{n+1/2}) - \tfrac{1}{2}k(A_f A^{-1} - I)g'(t_{n+1}).$$

In order to implement the split scheme, we also need V_1^* and V_1^{n+1}. We want $V_1^* = v^*(h, t_{n+1/2}) = v^*(0, t_{n+1/4})$ and so the appropriate value comes from the second equation of

$$\vec{u}^*(0, t_{n+1/4}) \approx g(t_{n+1/4}) + \tfrac{1}{4}k(A_f A^{-1} - I)g'(t_n)$$

i.e.,

$$V_1^* = g_2(t_{n+1/4}) + \frac{k}{4(2-\epsilon_1\epsilon_2)}(2\epsilon_2 g_1'(t_n) + \epsilon_1\epsilon_2 g_2'(t_n))$$

where g = $(g_1, g_2)^T$. Similarly,

$$V_1^{n+1} = g_2(t_{n+3/4}) - \frac{k}{4(2-\epsilon_1\epsilon_2)}(2\epsilon_2 g_1'(t_{n+1}) + \epsilon_1\epsilon_2 g_2'(t_{n+1})).$$

Computations confirm that these boundary conditions preserve $O(\epsilon k^2)$ global accuracy in the split scheme. Actually, for this particular example with $k = 2h$, even greater accuracy can be achieved. Computing $E_s(k)$ from (2.5a) shows that the $O(\epsilon k^3)$ terms exactly cancel the $O(\epsilon k^3)$ terms in $E_{\text{split}}(k)$, and that the total truncation error $E^{TSM}(k)u$ is actually $O(\epsilon^2 k^3)$, giving $O(\epsilon^2 k^2)$ global accuracy. Higher order boundary conditions can be derived which maintain this accuracy, but this cancellation of errors is a fluke which does not occur in general.

Stability for the initial boundary value problem. The boundary approximations derived here all depend only on the given boundary function $g(t)$ and its deri vatives. Suppose the time-split, method used in the interior is Cauchy stable. Then the stability of the resulting scheme for the initial boundary value problem follows directly from the theory of Gustafsson, Kreiss and Sundström[8], if we modify their stability definition 3.3 by using an appropriate Sobolev norm of the boundary data on the right- hand side.

8. Computational results.

In this section WC give various examples of splittings and present the results of some numerical experiments. The first example is a 2×2 upper triangular system of the form (1.11). We' demonstrate the effects of the splitting error and its reduction by the use of a simple change of variables as discussed in section 2.

The second example is a variable coefficient scalar equation in which the coefficient has small variations around some large mean value. We give an expression for the splitting error in such problems.

In example 8.3 we consider the one-dimensional shallow water equations. In some cases this system can be broken up into a constant fast part arid a quasilinear slow part in conservation form.

Example 8.1. This problem is designed to illustrate the effects of the splitting error. Consider

$$u_t = \begin{bmatrix} 10 & 1 \\ 0 & 1 \end{bmatrix} u_x \qquad \text{for } 0 \le x \le 1,\ t \ge 0 \tag{8.1}$$

with initial conditions

$$u_1(x,0) = u_2(x,0) = e^{-100(x-1/2)^2}$$

and periodic boundary conditions

$$u_j(0,t) = u_j(1,t) \qquad t \ge 0,\ \text{j} = 1,2.$$

Figure 8.1 shows the results after 236 time steps using Lax-Wendroff with $h = 1/50$ and $k = h/10$ on the unsplit problem. Figure 8.2 shows the results based on the splitting

$$A_f = \begin{bmatrix} 10 & 0 \\ 0 & 0 \end{bmatrix}, \quad A_s = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

We used $k = h = 1/50$ with

$$Q_s(k) = LW(A_s, k),\ Q_f(k/2) = (LW(A_f, k/10))^5.$$

In this case $E_s(k) = E_f(k/2) = 0$ by a judicious choice of k/h and m. The second component u_2 is computed exactly and the errors in u_1 are due entirely to the splitting error.

If the change of variables suggested in (2.18) is applied twice to (8.1) with $\epsilon = 0.1$, we obtain the new variable

$$\bar{u}_1 = u_1 - (\epsilon + \epsilon^2)u_2 = u_1 - 0.11212$$

and (8.1) becomes

$$\begin{bmatrix} \bar{u}_1 \\ u_2 \end{bmatrix}_t = \begin{bmatrix} 10 & 0.01 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \bar{u}_1 \\ u_2 \end{bmatrix}_x.$$

If we solve this system with the same split scheme as before and then transform back to the original variables by $u_1 = \bar{u}_1 + 0.1\,u_2$, the errors in u_1 are reduced to $O(10^{-3})$ as seen in figure 8.3.

The Leapfrog Duhamel scheme can be applied to this system with similar results. The same change of variables can clearly be used to reduce the splitting error in this scheme as well.

Example 8.2. For problems of the form

$$u_t = (a + \alpha(x))u_x$$

with a constant and $|\alpha(x)| << |a|$, the splitting error operator corresponding to $A_f = a$, $A_s = \alpha(x)$ is

$$\begin{aligned} E_{\text{split}}(k) &= \exp(\tfrac{1}{2}ka\partial_x)\exp(k\alpha(x)\partial_x)\exp(\tfrac{1}{2}ka\partial_x) - \exp(k(a+\alpha(x))\partial_x) \\ &= -\frac{1}{12}k^3a((\tfrac{1}{2}a + \alpha(x))\alpha''(x) - (\alpha'(x))^2)\partial_x + O(k^4). \end{aligned}$$

For the Leapfrog Duhamel scheme the splitting error is

$$\begin{aligned} &E_{\text{split}}(2k) + 2k\alpha(x)\partial_x E^I_{\text{split}}(k) \\ &\quad = E_{\text{split}}(2k) + 2k\alpha(x)\partial_x(\tfrac{1}{2}k^2a\alpha'(x)\partial_x + O(k^3)) \\ &\quad = -\frac{1}{3}k^3a[(2a + \alpha(x))\alpha''(x) - 4(\alpha'(x))^2 - 3\alpha(x)\alpha'(x)\partial_x]\partial_x + O(k^4). \end{aligned}$$

The Lax-Wendroff and leapfrog errors on $u_t = \alpha(x)u_x$ are respectively

$$\begin{aligned} E_s^{LW}(k) = &-\frac{1}{6}k^3\alpha(x)[\alpha^2(x)\partial_x^3 + 3\alpha(x)\alpha'(x)\partial_x^2 + ((\alpha'(x))^2 + \alpha(x)\alpha''(x))\partial_x] \\ &+ \frac{1}{6}kh^2\alpha(x)\partial_x^3 + O(k^4) \end{aligned}$$

and

$$E_s^{LF}(k) = 2E_s^{LW}(k) + O(k^4).$$

For the test problem

$$\begin{aligned} &u_t = (1 + 0.1\sin(2\pi x))u_x \qquad \text{on } [0, 1] \\ &u(x, 0) = \sin(4\pi x) \qquad 0 \le x \le 1 \\ &u(0, t) = u(1, t) \end{aligned}$$

a comparison of the errors shows that the splitting error for either scheme with $k = 4h$ should be of roughly the same size as $E_s(k)$ and considerably smaller than the error for the unsplit operator with the same spatial step and reduced time step $k = h/2$. Thus we expect the split scheme with the true solution operator used on $u_t = u_x$ to be more accurate than the unsplit scheme. This is confirmed by the computational results in Table 8.1. Note that in this case the improved accuracy was obtained using only about one eighth the work required for the unsplit scheme.

If Lax-Wendroff is used on the fast scale, $Q_f(k/2) = (LW(A_f, k/8))^4$, the corresponding error $2E_f(k/2)$ is roughly the same size as the error in the unsplit scheme. This error dominates in the resulting split scheme and so we get roughly the same accuracy as in the unsplit scheme. This is also illustrated in Table 8.1.

Example 8.3. The one-dimensional shallow water equations can be written as

$$\begin{bmatrix} v \\ \phi \end{bmatrix}_t = -\begin{bmatrix} v & 1 \\ \phi & v \end{bmatrix}\begin{bmatrix} v \\ \phi \end{bmatrix}_x \tag{8.2}$$

where $v(x, t)$ is the velocity and $\phi = gh$ with $h(x, t)$ the height and g the gravitational constant. Typically $\phi(x, t) = \hat{\phi} + \phi'(x, t)$ where $\hat{\phi}$ is constant and

$$|\phi'(x,t)| \ll |\hat{\phi}|$$
$$|v(x,t)| \ll |\hat{\phi}|.$$

With the change of variables

$$u(x,t) = \hat{\phi}^{1/2} v(x,t)$$

the system (8.2) becomes

$$\begin{bmatrix} u \\ \phi \end{bmatrix}_t = -\hat{\phi}^{-1/2} \begin{bmatrix} u & \hat{\phi} \\ \hat{\phi} + \phi' & u \end{bmatrix} \begin{bmatrix} u \\ \phi \end{bmatrix}_x.$$

The natural splitting is then

$$A_f = -\hat{\phi}^{1/2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$A_s(u, \phi) = -\hat{\phi}^{-1/2} \begin{bmatrix} u & 0 \\ \phi' & u \end{bmatrix}.$$

We have $\|A_s\| \ll \|A_f\|$. The matrix A_f is constant and the method of characteristics can easily be used for $Q_f(k/2)$. Furthermore, the problem on the slow scale can he written in conservation form. Since $\phi_x = \phi'_x$ WC have

$$A_s \begin{bmatrix} u \\ \phi \end{bmatrix}_t = \left(-\hat{\phi}^{-1/2} \begin{bmatrix} \frac{1}{2}u^2 \\ u\phi' \end{bmatrix} \right)_x$$

For the numerical experiments we used the initial conditions

$$u(x,0) = 0$$
$$\phi(x,0) = 16 + 0.1\sin(2\pi x) \qquad 0 \le x \le 1$$

and took $\hat{\phi} = 16$. We again used periodic boundary conditions and compared Lax-Wendroff on the unsplit problem with $k = h/20$ to the split scheme with $k = h$ on the slow scale and the method of characteristics for $Q_f(k/2)$. For $h = 1/100$ the results are shown in table 8.2. Again the split scheme outperforms the unsplit scheme. The errors were reduced by a factor of 100 while at the same time the work was reduced by roughly a factor of 10.

Acknowledgments. We wish to acknowledge Gunilla Sköllermo's participation in the early phase of this project. Her examples and comments helped to steer us in the right direction. Computer time was provided by the Stanford Linear Accelerator Center of the U.S. Department of Energy. The paper was produced using TEX, a computer typesetting system created by Donald Knuth at Stanford.

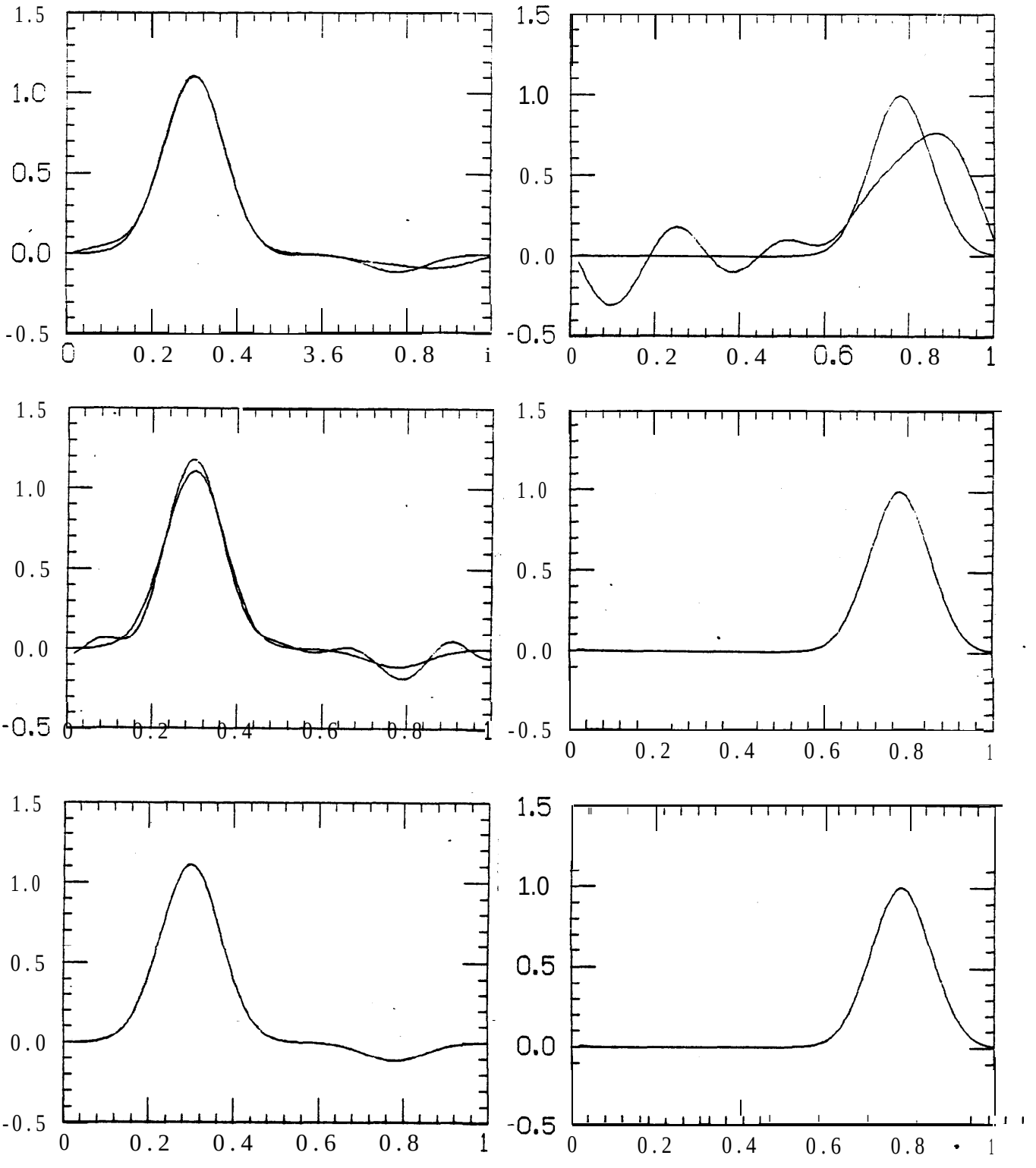

Figure 8.1. True and computed solutions at $t = 4.72$ for example 8.1. The first component, u_1, is on the left and the second component, u_2, is on the right. The schemes used are:

top: Unsplit Lax-Wendroff

middle: Time-split method (2.4a,b)

bottom: Time-split method with change of variables

Table 8.1. Max-norm errors for example 8.2 at various times t. The schemes used are:
#1: unsplit Lax-Wcndroff with $k = h/2$
#2: Leapfrog Duhamcl with $k = 4h$, $Q_f(k) = \exp(ka\partial_x)$
#3: Time-split method (2.4a,b) with $k = 4h$
#4: Time-split method (2.4a,c) with $k = 4h$, $m = 8$.

h	t	#1	#2	#3	#4
1/50	0.48	6.619(−2)	1.336(−2)	2.147(−3)	6.470(−2)
	0.96	1.342(−1)	1.949(−3)	4.598(−3)	1.315(−1)
	1.52	2.058(−1)	1.414(−2)	7.193(−3)	2.016(−1)
	2.00	2.685(−1)	3.434(−3)	9.617(−3)	2.623(−1)
1/100	0.48	1.677(−2)	3.356(−3)	5.581(−4)	1.635(−2)
	0.96	3.389(−2)	4.130(−4)	1.166(−3)	3.320(−2)
	1.52	5.314(−2)	3.365(−3)	1.845(−3)	5.197(−2)
	2.00	6.971(−1)	2.028(−4)	2.437(−3)	6.818(−2)

Table 8.2. Max-norm errors for u and ϕ in example 8.3 at various times t. The schemes used are:
#1: unsplit Lax-Wendroff with $h = 1/100$, $k = h/20$
#2: Time-split method (2.4a,b) with $k = h = 1/100$.

t	#1	#2
0.25	3.983(-4)	2.952(-6)
	3.354(-5)	2.338(-7)
0.50	8.059(−4)	5.882(−6)
	1.248(−4)	9.386(-7)
0.75	1.232(−3)	8.793(-6)
	2.683(−4)	2.085(-6)
1.0	1.687(−3)	1.166(−5)
	4.829(−4)	3.628(-6)

REFERENCES

[1] Abarbanel, S., and Gottlieb, D., Optimal time splitting for two and three dimensional Navier-Stokes equations with mixed derivatives, ICASE Report No. 80-6, 1980.

[2] Browning, G., Kasahara, A., and Kreiss, I-I.-O., Initialization of the prirnative equations by the bounded derivative method, J. Atmospheric Sci. 37(1980), pp. 1424-1436.

[3] Ccrtaine, J., The solution of ordinary differential equations with large time constants, in ***Mathematical Methods for Digital Computers,*** A. Ralston and I-I.S. Wilf, eds., Wiley, New York, 1960, pp 128- 132.

[4] Engquist, B., Gustafsson, B., and Vrecburg, J., Numerical solution of a PDE system describing a catalytic converter, J. Comp. Phys. 27(1978) pp. 295-314.

[5] Gadd, A.J., A split explicit integration scheme for numerical weather prediction, Quart. J. Royal Met. Soc. 104(1978), pp 569-582.

[6] Gourlay, A.R., Splitting methods for time dependent partial differential equations, in ***The State of the Art in Numerical Analysis,*** D. Jacobs, ed., Academic Press, 1977.

[7] Gustafsson, B., The convergence rate for difference approximations to mixed initial boundary value problems, Math. Comp. 29(1975) pp. 396-406.

[8] Gustafsson, B., Kreiss, I-I.-O., and Sundström, A., Stability theory of difference approximations for mixed initial boundary value problems. II, Math. Comp. 26(1972) pp. 649-685.

[9] Kreiss, IL-O., Problems on different time scales for ordinary differential equations, SIAM J. Numer. Anal. 16(1979) pp 980-998.

[10] Kreiss, H.-O., Problems with different time scales for partial differential equations, Comm. . Pure and Appl. Math. 33(1980), pp. 399-439.

[11] Lawson, J.D. and Morris, J.Ll., A review of splitting methods, Report CS-74-09, Department of Applied Analysis and Computer Science, University of Waterloo, 1974.

[12] Majda, G., Filtering techniques for oscillatory stiff ordinary differential equations, to appear.

[13] Mitchell, A.R., ***Computational Methods in Partial Differential Equations,*** Wiley, 1969,

[14] O'Malley, R.E., and Anderson, L.R., Singular perturbations, order reduction, and decoupling of large scale systems, in ***Numerical Analysis of Singular Perturbation Problems,*** Hemkes and Miller, eds., Academic Press, 1979, pp 317-998

[15] Richtmyer, R.D., Morton, K.W., ***Difference Methods for Initial- Value Problems,*** Interscience Tracts in Pure and Applied Math., No. 4, Wiley, 1967.

[16] Strang, G., On the construction and comparison of difference schemes. SIAM J. Numer. Anal. 5(1968) 506-517.

[17] Strikwerda, J., A time-split difference scheme for the compressible Navier-Stokes equations with applications to flows in slotted nozzles, ICASE Report No. 80-27, 1980.

[18] Turkel, E. and Zwas G., Explicit large time-step schemes for the shallow water equations, AICA Proceedings No. 3(1979) pp 65-69.