STABILITY OF FINITE DIFFERENCE BOUNDARY CONDITIONS

Jeff Thorson

Abstract

What are sufficient conditions on R to guarantee that the extrapolation of the differential operator $u_z = -RU$ is stable? There are two possible conditions of stability. The first (weak) condition is to require that at large enough z, the energy in u(z) is bounded by the initial energy of u(0). This is equivalent to the condition that all eigenvalues of R have a positive real part. It will be seen that in the case of an operator R with absorbing boundaries $u_- = ru_+$ ***it is sufficient that r lie in the upper right-hand quadrant of the complex plane***. The second stability condition is stronger: the energy in u(z) must be less than or equal to the initial energy u(0) for *all* z. For an operator with absorbing boundaries, this stronger condition is satisfied when, additionally, the following is known to be true:

$$\kappa(S) \leq e^{+mz}$$

where $\kappa(S)$ is the condition number of eigenvectors S of R, and m is the smallest real part among the eigenvalues of R. Normal operators that satisfy the weak stability criterion automatically are strongly stable. But this is not true for the class of non-normal operators, among which are R with absorbing side boundaries. Whether the above equation is satisfied or not depends on the particular operator R at hand. There are indications of what to do to enforce the above condition: one is to

increase the size of the attenuation factor ϵ. Interrelations between R and the second difference operator T that may simplify the task of satisfying the strong stability condition are given in the following sections of this paper.

Arbitrary Boundary Conditions

This discussion is restricted to operators in the frequency domain, that is, time has been transformed out. Generally, a one-way linear wave equation operator can be put in the form $u_z = -Ru$ in which the x-direction has already been discretized. R then is an n by n matrix and u is an n-vector, where n is the number of points in the x-direction of the differencing grid. The Crank-Nicolson approximation to this differential equation in z is

$$(I + \frac{\Delta z}{2} R)u = (I - \frac{\Delta z}{2} R)u_0 \qquad (1)$$

where u_0 is the initial condition (i.e. the solution at a previous level). Arbitrary boundary conditions specifying the relation between two points on either end of the grid are $u_0 = r_1 u_1$ and $u_{n+1} = r_n u_n$. Tacking these relations onto the matrix $A=(I + (\Delta z/2)\, R)$, which, say, is tridiagonal, we have the equivalent system

$$\begin{bmatrix} 1 & -r_1 & & & & & \\ \alpha_0 & \beta_1 & \alpha_2 & & & & \\ & \alpha_1 & \beta_2 & \alpha_3 & & & \\ & & x & x & x & & \\ & & & x & x & x & \\ & & & & \alpha_{n-1} & \beta_n & \alpha_{n+1} \\ & & & & & -r_n & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ \\ \vdots \\ \\ u_n \\ u_{n+1} \end{bmatrix} = \begin{bmatrix} 0 \\ d_1 \\ \\ \\ \\ d_n \\ 0 \end{bmatrix} \qquad (2)$$

For instance, when r_1, r_n equal 1, the boundary condition (B.C.) is zero-slope: $u_0 = u_1$. If $r_1, r_n = 0$, the B.C. is zero-value, where the zero value is attained midway between gridpoints 0 and 1 (respectively n and n+1). Now these end equations can always be absorbed into the adjacent equations by performing one step in the Gaussian elimination process. That is, note that system (3) is formed from (2) by subtracting a multiple of the first (and last) equation from the second (and next to last) equation.

$$\begin{bmatrix} (\beta_1+r_1\alpha_0) & \alpha_2 & & & & & \\ \alpha_1 & \beta_2 & \alpha_3 & & & & \\ & x & x & x & & & \\ & & x & x & x & & \\ & & & x & x & x & \\ & & & & \alpha_{n-1} & \beta_{n-1} & \alpha_n \\ & & & & & \alpha_{n-1} & (\beta_n+r_n\alpha_{n+1}) \end{bmatrix} \begin{bmatrix} u_1 \\ \\ \vdots \\ \\ u_n \end{bmatrix} = \begin{bmatrix} d_1 \\ \\ \vdots \\ \\ d_n \end{bmatrix} \quad (3)$$

In terms of the norm of u, (3) is the proper system to use, because the n-length vector u represents the actual energy in the grid rather than the (n+2)-length vector u in (2). This can be shown by re-deriving (3) in the following manner. Consider that the discrete system A is now infinite-dimensional. The banded matrix A continues without limit:

$$\begin{bmatrix} \ddots & \ddots & \ddots & & \\ \alpha_{k-1} & \beta_k & \alpha_{k+1} & & \\ & \alpha_k & \beta_{k+1} & \alpha_{k+2} & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} \vdots \\ u_{k-1} \\ u_k \\ u_{k+1} \\ u_{k+2} \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ d_{k-1} \\ d_k \\ d_{k+1} \\ d_{k+2} \\ \vdots \end{bmatrix} \quad (4)$$

Applying an internal boundary condition at point k of the grid is equivalent to "clamping" the solution there. The solutions on either side of this clamp then can proceed independently of each other. For example, the proper B.C. to "clamp" a tridiagonal operator at point k is $u_k = r u_{k+1}$. Adding the appropriate multiples of this equation to the k-th and (k+1)st equation in (4), it is possible to zero out the adjacent off-diagonal elements, which results in a decoupled system:

$$\left[\begin{array}{cccc|cccc} \ddots\ x & \ddots\ x & \ddots\ x & & & & & \\ & & \alpha_{k-1} & (\beta_k + r^{-1}\alpha_{k+1}) & 0 & & & \\ \hline & & & 0 & (\beta_k + r\alpha_{k-1}) & \alpha_{k+1} & & \\ & & & & x\ \ddots & x\ \ddots & x\ \ddots & \end{array}\right] \left[\begin{array}{c} \vdots \\ u_k \\ \hline u_{k+1} \\ \vdots \end{array}\right] = \left[\begin{array}{c} \vdots \\ d_k \\ \hline d_{k+1} \\ \vdots \end{array}\right]$$

(5)

As a practical matter, we must assume that $\alpha_{k-1} = \alpha_{k+1}$. It is interesting to note that the decoupled systems have the same boundary condition only when r is 1, 0, or -1. In the special case of absorbing sides, $\mathrm{Im}(r) > 0$ is known by experience to leak energy out while $\mathrm{Im}(r) < 0$ feeds energy in. The opposing B.C.'s in (5) exhibit opposite behavior -- $\mathrm{Im}(r^{-1}) < 0$ implies $\mathrm{Im}(r) > 0$, or one B.C. is stable while the other is unstable. More will be said later on what we mean by stability.

A similar clamping at the other end of the segment of interest yields the finite dimensional system (2). The norm of the solution vector u in (2) then naturally measures the energy in the segment $(u_1,\ldots,u_n)$. As a matter of fact, if r = 1 or -1 and α, β are all equal, the infinite dimensional problem is divided this way into a sequence of identical finite systems.

Let's proceed with the discussion using a concrete example: the 45-degree finite difference operator in the frequency domain (Jacobs et al., 1979). Assume constant velocity now. Later it will be allowed to vary. The operator, with the shift term removed, is

$$\frac{\partial u}{\partial z} = -Ru, \qquad R = \frac{1}{v}\left[\frac{v/\Delta x^2 T_0}{2s + \dfrac{v/\Delta x^2 T_0}{2s}}\right] \qquad (6)$$

where $s = \epsilon + i\omega$, v = velocity, and T_0, the second derivative differencing matrix, is defined to be

$$T_0 \triangleq \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & \ddots & \ddots & & \\ & & \ddots & \ddots & -1 & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix}$$

Expanding (6) out into the Crank-Nicolson form (1), yields

$$Au \equiv (\alpha I + (v + \beta)T_0)u = (\alpha I + (v - \beta)T_0)u_0 \equiv A_0u_0 \qquad (7)$$

where $\alpha = 4s^2\Delta x^2/v$ and $\beta = 2s$. Notice now that ***incorporating the B.C.*** $u_0 = r_1u_1$ ***into A is equivalent to incorporating it into*** T_0. That is,

$$\begin{bmatrix} 1 & -r_1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \end{bmatrix} = 0$$

added to $\begin{bmatrix} -(v + \beta) & \alpha + 2(v + \beta) & -(v + \beta) \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \end{bmatrix} = 0$

yields $\begin{bmatrix} 0 & \alpha + (2 - r_1)(v + \beta) & -(v + \beta) \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \end{bmatrix} = 0.$

So (7), with the general two-value boundary conditions $u_0 = r_1u_1$ and $u_{n+1} = r_nu_n$ added in, is equivalent to

$$(\alpha I + (v + \beta)T)u = (\alpha I + (v - \beta)T)u_0 \qquad (7a)$$

where now $T = T_0 + B$, that is,

$$T_0 \triangleq \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix} \qquad B \triangleq \begin{bmatrix} 1-r_1 & & \\ & \bigcirc & \\ & & 1-r_n \end{bmatrix} \tag{8}$$

Actually this trick can only be done for the 15-degree and 45-degree expansions of the wave equation operator. In higher order expansions (Godfrey et al., 1979), T^2 terms arise in (7). Later it will be seen, however, that putting the two-term boundary conditions, matrix B, into T in the higher order expansions produces a consistent higher order boundary in the Crank-Nicolson system (7). The manufactured B.C. will have the same effect as the original two-term B.C., e.g. it will continue absorbing.

To summarize this section, a two-term boundary condition (*a*) ***may be incorporated into adjacent equations to decouple a larger system into smaller systems, or*** (*b*) ***may be further incorporated, in the case of a 45-degree system, into T itself***. This indicates that if we assume that (7) is stable, the stability of the system (7a) does not primarily depend on the frequency-dependent coefficients α and β, but only on the value of r, the boundary coefficient. The stability of (7a) is examined in greater detail in the next section.

Stability Conditions for Non-normal Matrices

By stability we shall now mean the strong condition that if u satisfies the differential equation

$$u_z = -Ru \tag{9}$$

then $||u|| \le ||u_0||$ for ***any*** depth of propagation z below the initial wavefield u_0. An equivalent condition is that R satisfy $x^H(R + R^H)x \ge 0$ (Brown, 1979). Additionally, if (9) is stable in this sense, so is the Crank-Nicolson form (1) that is derived from (9). For a proof of this, see Godfrey et al. (1979). It should be mentioned that these conditions apply to solutions propagating downward (+z) and forward in time. Other directions necessitate various changes in sign in various places (Jacobs et al., 1979). Incidentally, throughout this paper the term ***positive real*** will mean that all eigenvalues of an operator have their real parts positive: $Re(\lambda_j) \ge 0$.

A ***normal*** matrix is one that commutes with its conjugate transpose, $NN^H = N^H N$. Normal operators are precisely those that have a complete set of orthonormal eigenvectors, so that N can be decomposed into $Q^H \Lambda Q$ where Q is unitary, and Λ is diagonal (Noble, 1969). It is easy to prove that ***given a normal R,*** $||u|| \le ||u_0||$ ***if and only if R is positive real, that is, the eigenvalues of R all have a positive real part.*** The proof goes as follows:

> Let us assume a slightly more general form for R. R may be non-normal, but assume that it still has a complete set of eigenvectors, i.e. it is not defective. Then R can be diagonalized to Λ by a similarity transformation (Noble, 1969).

$$R = S^{-1}\Lambda S$$

$$u_z = -Ru$$

Change coordinates, $w = Su, \quad u = S^{-1}w$

$$S^{-1}w_z = -RS^{-1}w$$

$$w_z = -SRS^{-1}w = -\Lambda w$$

The formal solutions w_j, given the initial values w_{0j}, are

$w_j = \exp(-\lambda_j \Delta z) w_{0j}$. Clearly the eigenvalues λ_j (in Λ) must all have their real parts greater than zero if w is to be bounded for all z. Therefore $||w|| \le ||w_0||$ for all z iff R is positive real. But

$$||u|| \le ||S^{-1}||*||w|| \le ||S^{-1}||*||w_0|| \le ||S^{-1}||*||S||*||u_0||.$$

But if R is ***normal***, then a unitary S exists, i.e. $||S|| = ||S^{-1}|| = 1$. Therefore, $||u|| \le ||u_0||$ iff R has positive real eigenvalues, which proves the assertion.

A stronger bound can be wrung out of the above proof. If m is the smallest real part among the eigenvalues λ_i,

$$||w|| \le ||w_0|| e^{-mz}$$

$$||u|| \le \kappa(S) ||u_0|| e^{-mz} \qquad (10)$$

where $\kappa(S)$ is the condition number of the matrix of eigenvectors S of R. The condition number is a measure of how far S is from orthonormality -- $\kappa(S) = 1$ if S is unitary. Now the inequality (10) holds for any R that is diagonalizable, which is a more general case than for a normal R.

We are now in a position to see that for reflecting boundary conditions (zero slope or zero value) the 45-degree equation (6) is strongly stable, because R shares common eigenvectors with T_0. T_0 in turn is positive real and is normal. It only remains to see that the eigenvalues of R are positive real -- it turns out that they are. See Brown (1979) for a proof.

Now, for more general boundary conditions, r in (8) may be complex, which makes $T = T_0 + B$ non-normal. In particular, this happens for Clayton and Engquist's absorbing boundary conditions (Clayton, 1977). Though the stronger condition $||u|| \le ||u_0||$ doesn't hold, it is encouraging to see in (10) that $||u||$ has an exponential bound, so even if the solution grows a bit at the beginning, it has to stabilize at large enough z. We can draw more inferences from the inequality (10): it is only in a particular range of z that $||u||$ can be greater than $||u_0||$, depending on a) how close to zero the smallest eigenvalue of R is, and b) how far the eigenvectors S of R deviate from orthogonality.

For wave equation migration, typically one applies (7) or (7a) in a large number of small steps of size Δz. If the operator doesn't change, (10) is still in effect, where z is the total distance migrated. Stability is guaranteed -- it is only a matter of going deep enough so that $\exp(-mz)\kappa(S) \leq 1$ is guaranteed. The vectors u at intermediate depths are somehow "conditioned" by repeated application of the operator so that they eventually stabilize.

Now if the operator *does* change at each step, for example in modeling a variable velocity medium, it is conceptually possible that u may grow without bounds. Suppose at the j-th z-level a new operator R_j is applied. The cumulative operator (matrizant)

$$M_j = \prod_{i=0}^{j} R_i$$

probably has eigenvectors that fluctuate about some average set of eigenvectors -- as long as $\kappa(S)$ is small -- and so (10) can be applied roughly to this average set. To restate, smoothly varying velocity is not foreseen to be a problem, provided that a) R is always positive real, and b) the non-normal boundary conditions applied are sensible enough so that $\kappa(S)$ is reasonably small.

There is a potentially more serious problem. If another operation is applied to u between downward continuation steps in z, it is possible that the "conditioning" of u may be effectively destroyed by this interposed operation. This may very well be done in a systematic manner so that the only bound that applies is the much weaker

$$||u_n|| \leq \left[\prod_{j=1}^{n} \kappa(S_j) e^{-m_j \Delta z} \right] ||u_0|| \qquad (11)$$

There are indications that this happens in shot-geophone migration, when the shot operator and geophone operator are alternately applied to the two-dimensional u field. The shot operator may upset the conditioning of u by the geophone operator at each step, and vice versa. A way to guarantee that this doesn't happen is to impose the condition

$$\kappa(S) \le e^{+m_j \Delta z} \tag{12}$$

at each step. This involves a knowledge of S_j and λ_j for each step, which is hard to come by. This will be examined later for the case of the 45-degree operator.

Boundary Conditions and the Expansions to the Square Root Operator

The monochromatic one-way wave equation operator R of (9) may be approximated by a continued fraction expansion (Godfrey et al., 1979).

$$R = \frac{1}{V}\left[\cfrac{VTV/\Delta x^2}{2s + \cfrac{VTV/\Delta x^2}{2s + \cfrac{VTV/\Delta x^2}{2s + \cdots}}}\right] \tag{13}$$

Here $s = \epsilon + i\omega$ and V is a diagonal matrix containing the velocity function v(x). VTV is now the second derivative operator with the general boundary conditions included:

$$VTV = \begin{bmatrix} (2-r_1)v_1^2 & -v_1v_2 & & & \\ -v_1v_2 & 2v_2^2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -v_{n-1}v_n \\ & & & -v_{n-1}v_n & (2-r_n)v_n^2 \end{bmatrix} \tag{14}$$

For the rest of the analysis we are going to ignore the 1/V term in front of the continued fraction (13), for reasons given in Brown (1979) and Godfrey (1979).

When the denominator of the continued fraction is premultiplied out, it yields a Crank-Nicolson form which is a generalization of (7a):

$$Mu = \left[\sum_{j=0}^{n} \alpha_j (VTV)^j\right] u = d \qquad (15)$$

The first criterion to meet is consistency, that is, whether the B.C. incorporated into VTV does the same thing as an equivalent B.C. imposed on the system (15). We have seen that this is trivially true for the first two terms of the expansion (13), since in that case (15) only goes up to the first power of VTV. Although it won't be proved, there is good indication that it is true for all higher orders in (15). For example, starting with VTV of the form (14), the boundary condition on $(VTV)^2$ can be shown to be

$$(\ 1 \quad -2 \quad 1\ ) * (\ 1 \quad -r_1) \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \end{bmatrix} = 0$$

where * represents a convolution. This is virtually the same as the boundary condition

$$(\ 1 \quad -r_1) \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 0$$

Now if all the powers of VTV in (15) have consistent B.C.'s, their sum certainly will. This seems to be a natural way to generate higher order B.C.'s for banded matrices broader than a tridiagonal.

With the consistency question out of the way (albeit crudely), the most important point with respect to stability is to ***guarantee that R is positive real***. This has been shown to be true for normal matrices R by

Brown (1979). The rest of this section will be concerned with deriving a sufficient condition that will guarantee R is positive real, ***even though it is not normal due to the boundary conditions imposed on VTV***. Namely, this condition is that VTV be both positive real and "positive imaginary" so to speak. Let's state this as a theorem. First, we know that R shares a common set of eigenvectors with VTV. Second, the eigenvalues of R are related to those of VTV by the relation (13) with λ_R, λ_{VTV} replacing R and VTV respectively. As in the previous section, T can be written as $T_0 + B$, where T_0 and B are given in (8).

Theorem: ***It is sufficient that*** $1-r_1$, $1-r_n$ ***lie in quadrant I of the complex plane for*** $u_z = -Ru$ ***to be stable***. ***That is,*** $Re(\lambda_R) \geq 0$ ***only if*** $Re(1-r_{1,n}) \geq 0$ ***and*** $Im(1-r_{1,n}) \geq 0$.

By stability we mean the weaker condition that R have positive real eigenvalues. The proof of this follows an induction argument parallel to that given in Brown (1979) and in Jacobs (1979):

> First assume $(1-r_1)$ and $(1-r_n)$ lie in quadrant I. The eigenvalues of VBV, which are $v_1^2(1-r_1)$, $v_n^2(1-r_n)$ and zero, all lie in quadrant I. Applying Muir's rule number 3 (Brown, 1979) to both $VTV = VT_0V + VBV$ ***and*** $VTV/i = (VT_0V + VBV)/i$, observe that the eigenvalues λ_{VTV} of VTV lie in quadrant I. Now (13) can be rewritten in terms of the eigenvalues as the recursion formula
>
> $$\lambda_k = \frac{1}{\dfrac{2s\Delta x^2}{\lambda_{VTV}} + \dfrac{1}{2s + \lambda_{k-2}}} \qquad \begin{array}{l} \text{for } k = 2,4,6,\ldots,n \\ \text{or } k = 3,5,7,\ldots,n \end{array} \qquad (16)$$
>
> Assume n is even. Now since $|\arg(\lambda_{VTV}) - \arg(2s)| \leq \pi/2$, $\lambda_0 = \lambda_{VTV}/2s$ has to be positive real. By induction, assume λ_{k-2} is positive real. Since 2s and $2s\Delta x^2/\lambda_{VTV}$ are positive real, and all combinations of sums and inverses of positive real numbers are positive real, it follows that λ_k is positive real. Note that no ***products*** were involved in the derivation, because it is ***not*** necessarily true that the product of positive real numbers remains positive real. We can also prove stability for odd-power expansions of

(16) by noting that

$$\lambda_1 = \frac{1}{\frac{2s\Delta x^2}{\lambda_{VTV}} + \frac{1}{2s}} \qquad (16a)$$

is positive real. Then the exact same induction argument can be used as in argument given above. This proves the theorem.

Absorbing Boundaries

This section will discuss qualitatively some stability bounds on the 45-degree equation with absorbing sides. Assume now that velocity is a function of z only.

A common absorbing side boundary used (here, at least) is one given by Clayton (1977) -- a version derived by combining the 15-degree equation with a B.C. that is inhomogeneous in the Crank-Nicholson formula (7). In the frequency domain, it is given as

$$r_k = \frac{1 + i\alpha\omega\Delta x/v_k}{1 - i\alpha\omega\Delta x/v_k}$$

where k = 1,n. Now r_k is always positive real, and is positive imaginary when $\alpha \geq 0$. By the previous theorem this ensures that R is positive real. R, however, is not normal, but still has a complete set of eigenvectors for any value of (real) α. Propagation can be unstable only during times when the condition (12) fails. Is this ever possible? There are some tunable parameters that are at our disposal to suppress the instability if it does happen to occur -- mainly the parameter ϵ. Also, to get a better bound of the type (10), it is possible to derive a variant of (10) that applies to the Crank-Nicolson system (1). It is

$$||u|| \leq \kappa(S) * \left|\left| \frac{1 - \frac{\Delta z}{2}\lambda_m}{1 + \frac{\Delta z}{2}\lambda_m} \right|\right| * ||u_0||$$

It can be seen that we would like to have the minimum eigenvalue λ_m as large as possible.

In the specific case of the 45-degree operator with constant velocity, (16) reduces to

$$\lambda_R = \frac{v/\Delta x}{\frac{2s\Delta x}{v\lambda_T} + \frac{v}{2s\Delta x}} \tag{17}$$

where λ_T represents each eigenvalue of T, and $s = \epsilon + i\omega$. For the midrange frequencies, the two terms in the denominator of (17) are roughly of the same size (assuming reasonable values for v and Δx), so eigenvalues of R are at their approximate maximum. When ω is small, the second term in the sum dominates, so that $\lambda_R = 2s$. It is important not to let λ_R drop to zero, but this is easily handled by setting $\epsilon > 0$ in the second term. When ω is large, or for very small λ_T, i.e. for low wavenumbers, the first term dominates. Then $\lambda_R = v^2\lambda_T/2s\Delta x^2$. Generally, the larger the dimension of our grid in the x-direction, the smaller the minimum λ_T becomes. The ϵ in the first term of the denominator of (17) has virtually no effect on the stability of R, compared to the advantageous effect it has in the second term. It seems then that the first term may be somewhat uncontrollable, and the only way around disturbances possibly generated at the boundaries by a large first term would be to restrict the frequency range or otherwise filter out low wavenumbers from u before operating.

REFERENCES

Brown, D., 1979, Muir's rules for matrices: Another look at stability: SEP - this volume.

Clayton, R. and B. Engquist, 1977, Absorbing boundary conditions: SEP-11, p. 116.

Godfrey, B., F. Muir and J.F. Claerbout, 1979, Stable extrapolation: SEP-16, p. 83.

Jacobs, B., B. Godfrey, F. Muir, J.F. Claerbout and J.P. Dut, 1979, Bullet-proofing the code for the 45-degree equation: SEP-16, p. 89.

Noble, B., 1969, Applied linear algebra: Prentice-Hall.