Introduction to Numerical Relativity

1 Error Analysis

Let consider a differential system of the form

$$Lu = s \ , \tag{1}$$

where L is any differential operator, u is the unknown function, and s is some specified function of the independent variables and of the unknown but not of its derivatives $s = s(u, :)$, frequently called *sources*.

For illustrative purposes it is convenient to focus on problems that have at most two independent variables. We can think of u as a single function of one space variable and time $u = u(x, t)$, but our discussion applies to cases with more independent variables as well as multiple dependent variables.

We will use a superscript h, the grid spacing, to indicate that the functions or operators are discrete, and are associated with the Finite Difference Approximation (FDA)[1]. With this notation, we denote the FDA of (1) as

$$L^h u^h = s^h \ . \tag{2}$$

Let us introduce some important terminology relative to the different errors that will appear in our discrete description.

Truncation error

The truncation error τ^h, of a FDA is defined as

$$\tau^h := L^h u - s^h \ , \tag{3}$$

where u satisfies the continuum PDE. Note that the form of the truncation error can always be computed from the finite difference approximation and the differential equation.

Order of the FDA

We say that the FDA is pth order if

$$\lim_{h\to 0} \tau^h = O(h^p) \qquad \text{for some integer } p \ . \tag{4}$$

Consistency

When the FDA operator in the continuous limit ($h \to 0$) approach the original differential operator, we say that our approximation is consistent. In other words an scheme is consistent in the truncation error tends to zero in the limit $h \to 0$. This is considered the minimum requirement for any given finite difference approximation when the grid is refined.

A consistent FDA reduces locally to the differential equation in the continuum limit. In practice however, one is interested in another property. What one really looks for is an approximation of the solution of the continuous problem and this is related to the following definition.

Solution error

[1] In general, space and time grid-spacing can be assumed to be proportional to each other

The solution error associated with a FDA is defined as the difference

$$e^h := u - u^h \ . \tag{5}$$

Convergence

We say that the discrete solution converges if and only if

$$u^h \to u \quad \text{as} \quad h \to 0. \tag{6}$$

In other words we ask that solution error must tend to zero.

Stability

There is another very important property of FDAs required to solve in a satisfactory way the continuous problem. Independently of the behaviour of the solution u, we must ask that the solution to the finite difference equations u^h, should remain bounded after a given finite time t for any time step. This requirement is known as stability. Basically we are asking that no component of the initial data should be amplified arbitrarily. Sometimes this property is written as

$$\|u^h\|_{t=\tau} \ \leq C_\tau \|u^h\|_{t=0} \ , \tag{7}$$

where C_τ is a constant independent of h and τ is an arbitrary finite time. Stability is a property of the FDA, and is essentially the discrete version of the definition of well posedness for a system of evolution equations.

Roundoff error

Most numerical analysis concerns the properties of solutions to finite difference equations. However, when attempting to solve such equations on a computer, all the computations are carried out only approximately because the numbers are stored with a finite precision.

Floating point representation is the method used by modern computers for storing real numbers. The set of mathematical operations performed on these approximations is called floating point arithmetic.

The error in the solution caused by the use of floating point arithmetic in solving a finite difference equation is called roundoff error.

Lax theorem

A very important result, sometimes dubbed the fundamental theorem of finite difference approximation is the Lax theorem which says: *Given a mathematical well posed initial value problem which is linear and a finite difference approximation to it that is consistent then the approximation is convergent if and only if is stable.* The importance of this theorem relies in the fact that stability and consistency are much easier to prove than convergence specially when we do not know the exact solution of the problem.

1.1 The advection equation

Let us consider the advection equation.

$$\partial_t u + a\partial_x u = 0 \ . \tag{8}$$

Given the initial data $u(x,0) = u_0(x)$, The solution is

$$u(x,t) = u_0(x - at) \ . \tag{9}$$

The importance of this equation relies in the fact that we can perform a rather complete treatment of the convergence of the approximation. Let us introduce the difference operators which are centred approximations of the exact partial derivatives.

$$D_x u_j^n := \frac{u_{j+1}^n - u_{j-1}^n}{2h} \ , \tag{10}$$

and

$$D_t u_j^n := \frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} \ , \tag{11}$$

which are second order in the grid spacing . The idea behind the error analysis is to consider the solution of the FDA as a continuum problem. We shall express the difference operators and the FDA solution as a series of h. For example, the series expansion of D_x and D_t are [2]

$$D_x = \partial_x + \frac{1}{6}h^2\partial_{xxx} + O(h^4) \ , \tag{12}$$

$$D_t = \partial_t + \frac{1}{6}\lambda^2 h^2 \partial_{ttt} + O(h^4) \ . \tag{13}$$

Then, we can calculate the different errors previously introduced.

The operator L for the advection equation in (1) is $(\partial_t + a\partial_x)$, the sources $s = 0$, $L^h = D_t + aD_x$ and $s^h = 0$. Then we have

$$(D_t + aD_x)u = \tau^h = \frac{1}{6}h^2(\lambda^2\partial_{ttt} + a\partial_{xxx})u = O(h^2) \tag{14}$$

This is the truncation error τ^h defined in (3), therefore L^h is second order $O(h^2)$.

Here we will introduce an important observation given by Richardson [3]. The solution of any FDA, u^h which uses a uniform mesh with fixed scale h and is centred, should have the following expansion in the limit of $h \to 0$

$$u^h(x,t) = u(x,t) + h^2 e_2(x,t) + h^4 e_4(x,t) + ... \tag{15}$$

where $u(x,t)$ is the continuum solution and e_i are continuum error functions which do not depend on h [4]. Richardson expansion is telling us what is the dependence of u^h with h. Given the Richardson expansion of u^h, we express the discrete system $L^h u^h$ as

$$L^h u^h = (D_t + aD_x)(u + h^2 e_2 + ...) \ , \tag{16}$$

and then

$$\left(\partial_t + \frac{1}{6}\alpha^2 h^2 \partial_{ttt} + a\partial_x + \frac{1}{6}ah^2\partial_{xxx} + ...\right)(u + h^2 e_2 + ...) = 0 \tag{17}$$

Then order by order in h we have

$$(\partial_t + a\partial_x)u = 0 \tag{18}$$

which shows the consistency of the difference approximation. Furthermore, for h^2

$$(\partial_t + a\partial_x)e_2 = \frac{1}{6}(a\partial_{xxx} + \alpha^2\partial_{ttt})u \tag{19}$$

Notice that this equation has the same nature than the original PDE.

Convergence test

Starting from the Richardson expansion, and computing finite difference solutions with different values of h, we can learn a great deal about the error in our approximations. The whole procedure of investigating the manner in which a particular FDA converges is known as convergence testing.

It is important to notice that there are no hard and fast rules for convergence testing; rather, one tends to tailor the tests to the specifics of the problem at hand, and one gains intuition as one works through more and more problems.

[2] We consider $\Delta x = h, \Delta t = \lambda h$.

[3] Richtmayer and Morton, *Difference methods for initial value problems* Interscience Publishers, NY. 1967.

[4] When the FDA is not completely centred, also odd powers of h appear in the expansion. We will use centred difference for simplicity.

A simple example of a convergence test that is often used in practice is the following:

Let us compute three distinct solutions u^h, u^{2h} , u^{4h} at resolutions h, $2h$ and $4h$ respectively using the same initial data [5]. We also assume that the finite difference meshes belong to each other, i.e. that the $4h$ grid points are a subset of the $2h$ points which are a subset of the h. In this way specific grid function values can be compared directly to one another. In some practical cases however this is not an easy task.

From the Richardson ansatz we expect:

$$u^h = u + h^2 e_2 + h^4 e_4 + ... \ , \tag{20}$$

and

$$u^{2h} = u + (2h)^2 e_2 + (2h)^4 e_4 + ... \ , \tag{21}$$
$$u^{4h} = u + (4h)^2 e_2 + (4h)^4 e_4 + ... \ . \tag{22}$$

We define the convergence factor $Q(t)$ as

$$Q(t) := \frac{\|u^{4h} - u^{2h}\|}{\|u^{2h} - u^h\|} \ , \tag{23}$$

the norm used in (23) could be any suitable spatial norm, for instance the discrete ℓ_2 norm:

$$\|u^h\|_2 = \left(\frac{1}{N} \sum_{j=1}^{N} (u_j^h)^2 \right)^{1/2} \ . \tag{24}$$

In a second order scheme, we will have

$$\lim_{h \to 0} Q(t) = 4 \ . \tag{25}$$

When the solution is convergent $Q(t)$ tends to a constant as h approaches to zero.

If we expect that our solution is p order accurate, then

$$u^h = u + h^p e_p + h^{p+2} e_{p+2} + ... \tag{26}$$
$$u^{2h} = u + (2h)^p e_p + (2h)^{p+2} e_{p+2} + ... \tag{27}$$
$$u^{4h} = u + (2^2 h)^p e_p + (2^2 h)^{p+2} e_{p+2} + ... \tag{28}$$
$$\tag{29}$$

and the convergence factor should be

$$Q(t) = \frac{\|u^{4h} - u^{2h}\|}{\|u^{2h} - u^h\|} \tag{30}$$
$$= \frac{\|2^{2p} h^p e_p - 2^p h^p e_p + ...\|}{\|2^p h^p e_p - h^p e_p + ...\|} \tag{31}$$
$$= 2^p + O(h^{p+2}) \ , \tag{32}$$

In practice, we can use additional levels of discretization, $8h$, $16h$, etc.. to extend this test to look for trends in $Q(t)$ to convince oneself that the solution is convergent.

Magnitude of the error

[5] The fact that we use even multiples of h is just matter of convenience.

A point-like subtraction of any two solutions computed with different resolutions, provide an estimate of the level of error. For example if we use u^h and u^{2h} according to the Richardson expansion we have

$$u^{2h} - u^h = (u + (2h)^2 e_2 + ..) - (u + h^2 e_2 + ...) = 3h^2 e_2 + O(h^4) \sim 3e^h \tag{33}$$

from where we can say that the error in the numerical solution e^h is about one third of the difference between the solutions with two resolutions. Of course, this is only valid when we have reached the convergence regime, i.e.. where the Richardson expansion is valid.

Richardson extrapolation

Sometimes it is possible to improve the accuracy of the numerical solution once we have two solutions calculated in two different grids. The key idea is to eliminate the leading order terms in the error of both solutions. Let us consider again u^h and u^{2h} and the linear combination

$$\tilde{u}^h = \frac{4u^h - u^{2h}}{3} \ , \tag{34}$$

We know that u^h and u^{2h} are second order accurate, but it turns that

$$\begin{aligned} \tilde{u}^h &= \frac{4(u + h^2 e_2 + h^4 e_4 + ...) - (u + 4h^2 e_2 + 16h^4 e_4 + ...)}{3} & (35) \\ &= u - 4h^4 e_4 + O(h^6) & (36) \end{aligned}$$

i.e $\tilde{u}^h$ is fourth order accurate [6]. This improvement should be taken with care because only works when we have started with a fairly accurate solutions and this is not always the case in true numerical calculations.

Independent residual method

So far, we have established that our numerical solution u^h is converging as $h \to 0$ however, how do we know we are converging to the solution of the original continuum problem u?

To ensure that the solution of the given set of equations is the correct one, we can use an independent residual. The independent residual method relies on solving the equations using two different methods of discretization. First we have to use one method and keep the solution, then discretize the same equations using a second independent technique, and look at the error left when applying the second method to the solution obtained from the first. The error obtained by this procedure is referred to as the residual.

Let us assume we have obtained a discrete solution to the discrete problem

$$L^h u^h = 0. \tag{37}$$

We have tested, by computing the convergence factor, that u^h converges as h tends to zero. This does not mean that our implementation of L^h is correct [7]. In order to test the implementation what one should consider a distinct discretization operator $\tilde{L}^h$. When we expand it

$$\tilde{L}^h = L + h^2 \tilde{e}_2 + h^4 \tilde{e}_4 + ... \tag{38}$$

and apply the operator $\tilde{L}^h$ on the solution u^h. If u^h is second order convergent to the solution u then

$$u^h = u + h^2 e_2 + O(h^4) \ , \tag{39}$$

and then

$$\begin{aligned} \tilde{L}^h u^h &= (L + h^2 \tilde{e}_2 + O(h^4) + ...)(u + h^2 e_2 + O(h^4) + ...) & (40) \\ &= Lu + h^2(\tilde{e}_2 u + L e_2) & (41) \\ &\sim O(h^2) & (42) \end{aligned}$$

[6] Now should be evident why we choose the particular combination of u^h and u^{2h} in (34).

[7] This may be because a typo while writing the code in the computer or in any of the algebraic routines we used in the code.

i.e., the quantity $\tilde{L}^h u^h$ will converge quadratically as tends h to zero. It is important to notice that this is not always the case. Although our first solution u^h seems convergent, in the sense that $u^{2h} - u^h \to 0$ as $h \to 0$ we might have something like

$$u^h = u + e_0 + he_1 + h^2 e_2 + ... \ , \tag{43}$$

in this case

$$\begin{aligned} \tilde{L}^h u^h &= (L + h^2\tilde{e}_2 + O(h^4) + ...)(u + e_0 + he_1 + h^2 e_2 + O(h^4) + ...) && (44)\\ &= Lu + Le_0 + hLe_1 + O(h^2) && (45)\\ &= Le_0 + hLe_1 + O(h^2) \sim O(1) \ , && (46) \end{aligned}$$

i.e. we will not see the expected convergence at all. As h tends to zero this expression tends to a constant, a clear indicator that something is not working properly in the implementation.

One has to check the rate of convergence of the residual to zero, but this depends on the discretization methods used. For example, if we use a second order scheme to obtain the solution, and check our solution using a second order independent technique, the rate of convergence of the residual will be second order. However if in the second step we use a first order scheme, then the residual will be only first order convergent. If the solutions do not converge at all, then there is a problem with the implementation and we can not trust the result.