FINITE DIFFERENCES

OSCAR REULA

1. Introduction

Finite differences are linear operators (matrices) that act on finite arrays and try to approximate derivative operators acting on functions, that is, on infinite dimensional vectors. So the first step is to represent the functions as finite vector by using some of its values, as distributed on a finite array on the domain of dependence. In the simplest case that array consists of evenly spaced points in the domain, this might not the most efficient nor the most practical method, but it is the simplest.

Thus we take an array $\{x_i = x_0 + i * dx\}$, $i = 0, \dots N-1$, $dx = \frac{x_f - x_0}{N-1}$ of grid points, and we denote by $u_i := u(x_i)$ the image of u by the grid map.

2. The simplest finite difference operators approximating $\frac{d}{dx}$

We define:

$$(D_\pm u)_i := \frac{\pm(u_{i\pm1} - u_i)}{dx}$$

as the approximations to the limits to the left (right) of the derivative definition.

How big is the error we make with these approximations? We can answer this by assuming u is smooth and using Taylor's series, Expanding u we get,

$$u_{i\pm1} = u_i \pm \frac{du}{dx}|_{x_i} dx + \frac{d^2u}{dx^2}|_{x_i} \frac{dx^2}{2} \pm \frac{d^3u}{dx^3}|_{x_i} \frac{dx^3}{6} + \dots,$$

and so,

$$(D_\pm u)_i := \frac{\pm(u_{i\pm1} - u_i)}{dx} = \frac{du}{dx}|_{x_i} \pm \frac{d^2u}{dx^2}|_{x_i} \frac{dx}{2} + O(dx^2).$$

Thus, if our solutions are smooth, and dx is small enough, then these differences are good approximations. But it is easy to get a much better approximation by exploiting the fact that terms come with alternating signs, by taking the average of these two approximations we find a better one:

$$(D_0 u)_i := \frac{(D_+u)_i + (D_-u)_i}{2} = \frac{du}{dx}|_{x_i} + \frac{d^3u}{dx^3}|_{x_i} \frac{dx^2}{6} + O(dx^3)$$

This operador, called the center finite difference operator is a much convenient one, not only for its precision, but also –as we shortly see– for the properties of its eigenvalues.

Notice that if we apply these operators to a polynomial x^n we get that the first two are exact for $n = 0, 1$ (for all higher order derivatives in the error expression vanish), while the centered operator is exact even for the $n = 2$ case. So, one easy way to check the accuracy of the operators, or even to find them, is to apply them to polynomials of growing order.

3. Higher accuracy operators

It is easy to find finite difference operators with smaller errors, for instance, to find the centered finite difference operator of 4^{th} order we define it with free coefficients and impose the accuracy, either via Taylor's expansions or using polynomials,

$$(D_4 u)_i := \frac{a u_{i+2} + b u_{i+1} - b u_{i-1} - a u_{i-2}}{dx}$$

We have taken the symmetric case for it already guarantees that all even polynomials centered at x_i would vanish. Thus we need to try with the first and third order polynomials. These two conditions fix the values of a and b. So we apply it to $u_i = i * dx$, and $u_i = i^3 * dx^3$. In the first case, since the derivative is 1 we would need to get also 1, while in the second we would impose it to vanish.

$$(D_4(i*dx))_i = a2 + b + b + a2 = 1 \tag{1}$$

$$(D_4(i*dx)^3)_i = a2^3 + b + b + a2^3 = 0 \tag{2}$$

From where we see that $(a = \frac{-1}{12}, b = \frac{2}{3})$

$$(D_4 u)_i := \frac{-u_{i+2} + 8u_{i+1} - 8u_{i-1} + u_{i-2}}{12dx}$$

4. Finite Difference Operators in the Circle

So far we have considered local approximations to the derivative operator, we now specialize to the case where the domain is a circle of length L. That is we consider periodic functions of period L. For this case there are no limitations at the beginning or ends of our grid, for we take extra values using the periodicity. That is, for instance $x_N = x_0$, $x_{N+1} = x_1$, and so on. In particular $u^N = u^0$, etc. Thus, we shall consider a discrete vector of the form: $x_i = x_0 + i * dx$, $u_i = u(x_i)$, $i = 0, \dots N-1$, $dx = \frac{x_f - x_0}{N}$.

In that case the matrix representation of the operators is as follows:

$$D_+ = \frac{1}{dx}\begin{pmatrix} -1 & 1 & 0 & \dots \\ 0 & -1 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & -1 & 1 \\ 1 & \dots & 0 & -1 \end{pmatrix}$$

$$D_0 = \frac{1}{2dx}\begin{pmatrix} 0 & 1 & 0 & \dots & -1 \\ -1 & 0 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & -1 & 0 & 1 \\ 1 & 0 & \dots & -1 & 0 \end{pmatrix}$$

Two aspects of these matrices come up front immediately, first they are sparse, so applying them should not be too costly. Second that while the one for D_0 is antisymmetric, the one for D_+ is not. The antisymmetry of the second would tell us immediately that is diagonizable and its eigenvalues are imaginary. This has important consequences on stability.

4.1. **Eigenvectors/Eigenvalues of finite difference operators.** We now look at the eigenvalue-eigenvector pairs for the finite difference operators in the circle of length L.

Recall that for the derivative the eigenfunctions are:

$$\frac{d}{dx}e^{i2\pi kx} = i2\pi k e^{i2\pi kx},$$

periodicity would tell that we can only consider k's so that $i2\pi kL = i2\pi n$, for some arbitrary integer n, that is, $kL = n$, that is an infinite set of $k_n = \frac{n}{L}$. But since we are only considering grid points there immediately comes an extra condition, indeed, since $x_l = dx * l = \frac{L}{N} * l$ (for the periodic case we have $dx = \frac{L}{N}$ and not $\frac{L}{N-1}$), we have,

$$i2\pi k_n x_l = i2\pi \frac{n}{L}\frac{L}{N}l = i2\pi \frac{n}{N}l$$

Thus, the action of k_n is the same as the one of k_{n+N} and so we effectibly are limited to finite wave numbers,

$$k_n = \frac{n}{L}, \quad -N/2 \leq n \leq N/2, \quad N \text{ even}$$

At the discrete level we then have a basis $\{u_n^l\} = \{e^{i2\pi k_n x_l}\} = \{e^{i2\pi \frac{n}{N}l}\}$ of $n = -N/2, \ldots, N/2$, $(N+1)$ vectors each of $l = 0, \ldots N$, $(N+1)$ components each.

Let us now apply our finite difference operators to them,

$$(D_\pm e^{i2\pi k_n x})_l = \pm\frac{e^{i2\pi k_n x_{l\pm1}} - e^{i2\pi k_n x_l}}{dx} = \pm\frac{e^{i2\pi k_n x_l}(e^{\pm i2\pi k_n dx} - 1)}{dx} = \pm\frac{e^{\pm i2\pi k_n dx} - 1}{dx}e^{i2\pi k_n x_l}$$

where we have used that $e^{i2\pi k_n x_{l\pm1}} = e^{i2\pi k_n (x_l \pm dx)} = e^{i2\pi k_n x_l}e^{\pm i2\pi k_n dx}$.

Thus, we see that these functions are all eigenvectors of our operators. Notice that they are complex, in particular with real parts.

Things get much better if we look at the eigenvalues of D_0, from its definition we see that,

$$\begin{aligned}(D_0 e^{i2\pi k_n x})_l &= \frac{(e^{i2\pi k_n dx} - 1) - (e^{-i2\pi k_n dx} - 1)}{2dx}\, e^{i2\pi k_n x_l} \\ &= \frac{e^{i2\pi k_n dx} - e^{-i2\pi k_n dx}}{2dx}\, e^{i2\pi k_n x_l} \\ &= \frac{i\sin(2\pi k_n dx)}{dx}\, e^{i2\pi k_n x_l}\end{aligned}$$

Exercise: Compute the eigenvalues for the fourth oder operator defined above.

The above figure shows the dispersion relations for the different finite difference approximations up to order dx^8. Notice that they approach the linear dispersion relation of the operator $\frac{d}{dx}$ for small wave numbers, but that they depart for larger values. In fact, they go to zero for the larger grid wave numbers.

4.2. **Fourier Interpolation.** For numerical approximations one does not uses Fourier Series, but rather Fourier Interpolation. This is natural for we do not have at our disposal the values of functions at all points, but rather at a finite number of them.

Consider any periodic function $f(x)$ in $[0, L]$, and a grid, that for convenience we shall take to have an even number of points, $x_l = dx * l$, $l = 0, \ldots, 2M$, $dx = \frac{L}{2M+1}$. Let $f_l = f(x_l)$ be the restriction of $f(x)$ to the grid. Can we represent it as a linear combination of our base vectors, $\{\mathbf{e}_n := e^{i2\pi k_n x_l}\}$? Namely, does there exist another vector $\tilde{f}_m$ such that,

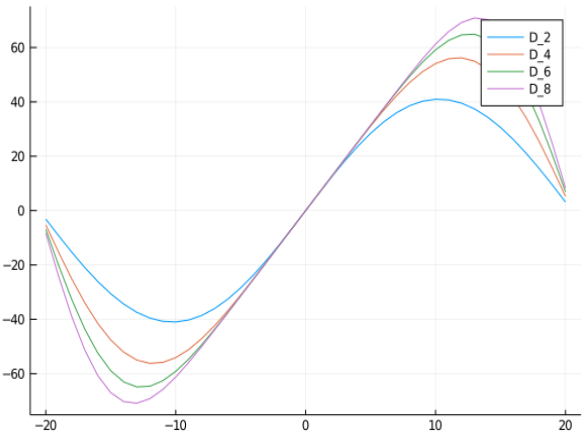

FIGURE 1. Dispersion relations for the different finite difference approximations.

$$f_l := f(x_l) = \sum_{m=-M}^{M} \tilde{f}_m e^{i2\pi k_m x_l}.$$

The answer is rather trivial, since they are eigenvectors of an anti-hermitian matrix they expand the space, so, as a matrix, its determinant is non-vanishing. Even more, they are ortogonal among each other, so, we can invert it easily. Notice that,

$$\frac{1}{L}\sum_{l=0}^{2M} e^{-i2\pi k_m x_l} e^{i2\pi k_n x_l}\, dx = \begin{cases} 1 & n = m \\ 0 & n \neq m \end{cases}$$

Where we have used that,

$$\sum_{l=0}^{2M} 1 = 2M + 1 = \frac{L}{dx},$$

and,

$$\begin{aligned} \sum_{l=0}^{2M} e^{-i2\pi(k_m - k_n)x_l} &= \sum_{l=0}^{2M} e^{-i2\pi \frac{(m-n)}{2M} l} \\ &= \sum_{l=0}^{2M} (e^{-i2\pi \frac{(m-n)}{2M}})^l \\ &= \frac{1 - (e^{-i2\pi \frac{(m-n)}{2M+1}})^{2M+1}}{1 - (e^{-i2\pi \frac{(m-n)}{2M+1}})} \\ &= 0, \quad n \neq m \end{aligned}$$

The actual implementation of this sum, or the inverse can be done much more efficiently that the naive counting of $\mathcal{O}(M^3)$, in fact it is only order $\mathcal{O}(Mlog(M))$. And it is called the Fast-FourierTransform. This efficiency makes possible to use the FFT for computing approximations to derivatives with exponential convergence (when the functions are smooth). The process is:

$$\{f_i\} \;\to\; \{\tilde{f}_m\} \;\to\; \{i2\pi k_m \tilde{f}_m\} \;\to\; \{Df_i\},$$

and comes with the name of *Spectral Methods*.

5. The advection equation.

We are now going to solve for an approximate solution to the advection equation, using our finite difference operators. We consider the equation,

$$\partial_t u = a\partial_x u, \qquad x \in [0, L]$$

with a real and positive. Given an initial data, $u_0 = f(x)$ the solution is, $u(t, x) = f(at + x)$, so the solution is a shift to the left with speed $-a$.

5.1. **The semi-discrete approximation.** We are now going to solve for a semi-discrete approximation,

$$\partial_t v_i = a(D_p v)_i$$

where $\{v_i(t)\}$ is a grid vector, and D_p is one of our order p finite-difference approximations to the space derivative.

Since D_p is anti-hermitian, it is diagonalizable and so there exists a base where D_p is diagonal and all its eigenvalues are pure imaginary. Thus we can expand v as,

$$v = \sum_n w_n \mathbf{e}_n \qquad \text{i.e.} \qquad v_l = \sum_n w_n e^{i2\pi k_n x_l}, \quad k_n = \frac{n}{L},$$

Thus, the coefficients of v in that base satisfy,

$$\partial_t w_n = a\lambda_n w_n$$

For instance, if we are using the second order centered finite difference operator, and taking as initial data $u_0(x) = w_n^0 e^{i2\pi k_n x}$, for some $k_n = n/L$, which corresponds to a grid vector $v_{nl} = e^{i2\pi k_n x_l} w_n$ then we would have,

$$\partial_t w_n = a\frac{i\sin(2\pi k_n dx)}{dx} w_n$$

Whose solution is,

$$w_n(t) = e^{ia\frac{\sin(2\pi k_n dx)}{dx}t} w_n^0$$

$$v_{nl}(t) = e^{ia\frac{\sin(2\pi k_n dx)}{dx}t + i2\pi k_n x_l} w_n^0$$

Corresponding to the continuum function $u_n(t, x) = e^{i2\pi k_n (a\frac{\sin(2\pi k_n dx)}{2\pi k_n dx}t + x)} w_n^0$. Thus, corresponding to a wave traveling with speed $a\frac{\sin(2\pi k_n dx)}{2\pi k_n dx}$. Notice that when $dx \to 0$ keeping k_n fixed, the speed tends to a.

Notice also that their group velocities, $\frac{d\lambda(k)}{dk}$ are negative! So that high frequency noice travels into the oposite direction!

$$\frac{d\lambda(k)}{dk} = a\frac{cos(2\pi k_n dx)}{2\pi dx}$$

Once we understand the propagation of each individual mode we can decompose our initial data into a sum of these modes and solve for them, resuming them at the end. This is very similar to do Fourier Analysis, but in the discrete is a bit different and comes under the name of Fourier Interpolation.

When dealing with a wave packet, there is interference and collective enhancements, in particular an important, and far from obvious, result is that their maximum propagation speed is larger than the exact one and of opposite sign. Thus, hight frequency modes are not at all well represented, travel into the opposite direction and with much larger speeds. We shall see these phenomena in the numerical examples.

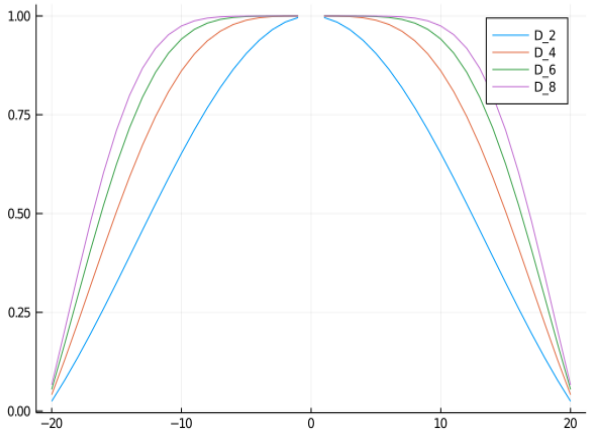

FIGURE 2. Face velocities for the different approximations.

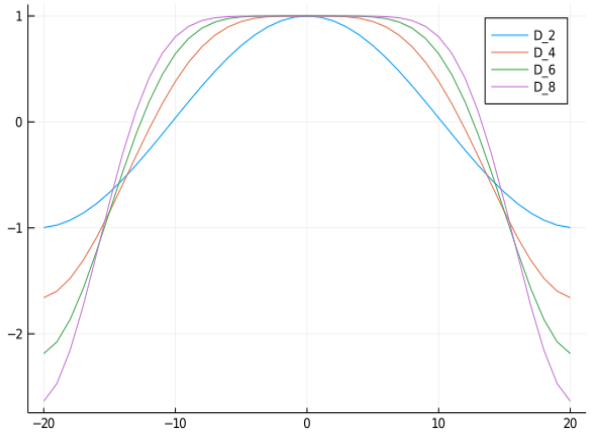

FIGURE 3. Group velocities for the different approximations.

5.2. **The continuum estimate.** We shall look now at the continuity of the solution with respect to the initial data in the Energy norms. We stay in the circle of length L. So we look at periodic functions $f(x+L) = f(x)$.

We define,

$$\mathcal{E}(t) := \frac{1}{2}\int_0^L u^2(t,x)\,dx,$$

and show that its value is bounded for all times, in fact, for this simple case, it is constant. Indeed, taking a time derivative of it we get,

$$\begin{aligned}\frac{d\mathcal{E}}{dt} &= \int_0^L u\partial_t u\,dx \\ &= \int_0^L ua\partial_x u\,dx \\ &= \frac{a}{2}\int_0^L \partial_x(u^2)\,dx \\ &= \frac{a}{2}[u^2|_L - u^2|_0] = 0.\end{aligned}$$

Therefore:

$$||u(t,\cdot)||^2_{L^2} = \mathcal{E}(t) = \mathcal{E}(0) = ||f(\cdot)||^2_{L^2}$$

This inequality (which for this simple case is an equality) is the key ingredient to understand the stability of hyperbolic equations: the solution depends continuously on the initial data. We see that the L^2 norm of the solution is preserved along the time interval. It is clear then that a Cauchy sequence of smooth initial data in L^2 would converge into an element of L^2 for each time t. Is that limiting solution smooth so that the equation applies?

Since $u_1 := u_x$ and $u_2 := u_{xx}$, satisfy the same equation we conclude that:

$$||u(t,\cdot)||^2_{H^2} = ||u(t,\cdot)||^2_{L^2} + ||u_x(t,\cdot)||^2_{L^2} + ||u_{xx}(t,\cdot)||^2_{L^2} = ||f(\cdot)||^2_{H^2}$$

but $H^2 \subset C^1$ from the Sobolev theorem, so we see that if the initial data $f(x)$ is in H^2 then the limiting solution is in C^1 and so a classical solution. We shall mimic this for the semi-discrete approximation.

5.3. **The semi-discrete estimate.** When considering the semi-discrete approximation we also have a similar estimate, indeed, we can define,

$$\mathcal{E}_D(t) := \frac{1}{2}\sum_{l=1}^N v_i^2\,dx,$$

and so,

$$\begin{aligned}\frac{d\mathcal{E}_D}{dt} &= \sum_{l=1}^N v_l\partial_t v_l\,dx \\ &= \sum_{l=1}^N v_l aDv_l\,dx \\ &= \frac{a}{dx}\sum_{l=1}^N\sum_{j=1}^N v_l A_{lj}v_j\,dx \\ &= 0.\end{aligned}$$

which vanishes since A is antisymmetric. Therefore:

$$||v(t)||^2_{l^2} = \mathcal{E}_D(t) = \mathcal{E}_D(0) = ||f(\cdot)||^2_{l^2}$$

6. The Initial Boundary Value Problem

If we try to solve the Advection problem,

$$\partial_t u = a\partial_x u, \qquad x \in [0, L], \quad a > 0$$

but this time in a line segment, that is without periodic boundary conditions, then, since we know the solution is a wave traveling to the left, we shall need to give boundary data at $x = L$. Thus, to solve the problem we not only give initial data but boundary data such that:

$$u(0, x) = f(x) \qquad u(t, L) = g(t)$$

In the continuum we have a proof of existence and uniqueness of solutions using the energy methods, which, for this simple case amounts to see that the energy, given by

$$\mathcal{E}(t) := \frac{1}{2}\int_0^L u^2(t, x)\, dx$$

is bounded for all times. Indeed, taking a time derivative of it we get,

$$\begin{aligned} \frac{d\mathcal{E}}{dt} &= \int_0^L u\partial_t u\, dx \\ &= \int_0^L ua\partial_x u\, dx \\ &= \frac{a}{2}\int_0^L \partial_x(u^2)\, dx \\ &= \frac{a}{2}[u^2|_L - u^2|_0] = 0. \end{aligned}$$

Where in the last line we have used.

Thus, if $a > 0$ then, integrating in time we get,

$$\mathcal{E}(T) \leq \mathcal{E}(0) + \frac{1}{2}\int_0^t au^2|_L\, dt = \frac{1}{2}[\int_0^L f^2(x)\, dx + \int_0^T ag^2(t)\, dt],$$

The numerical approximation must satisfy similar inequalities to preserve the corresponding stability. This is so if we requiere for the finite difference approximations to satisfy the Summation By Parts (SBP) property, namely,

$$\sum_{i,j=0}^{N-1} h^{ij}[u_i(Dv)_j + (Du)_i v_j] = u_{N-1}v_{N-1} - u_0 v_0,$$

for any pair of vectors u, v. This is the analogue to integration by parts, which is actually what we use in the energy inequality deduced above. Notice that in the sum which is our approximation for the integral we use weights which in general are not just Δx. This is so because the condition is not only on D, but rather on the pair (D, h).

For instance, for the case of the operator to second order accuracy, the pair is given by:

$$Du = \begin{cases} (Du)_0 &= \frac{u_1 - u_0}{\Delta x} \\ (Du)_i &= \frac{u_{i+1} - u_{i-1}}{2\Delta x} \qquad 0 < i < N-1 \\ (Du)_{N-1} &= \frac{u_{N-1} - u_{N-2}}{\Delta x}, \end{cases}$$

while,

$$h = \frac{1}{\Delta x} diag(\frac{1}{2}, 1, \cdots, 1, \frac{1}{2}).$$

Exercise: Check the SBP property for this operator.

There exist higher precision operators with the SBP property, for second and fourth order they are unique, but starting with the sixth order, there is freedom in how to choose them. There are several parameter families of them. From the second order case above it is clear that the precision drops to first order at the boundary. This is general, there is a drop of one order at boundaries, not only at the last point, but also at a finite number of points. Nevertheless the overall accuracy, as measured en the norm given by the scalar product defined by h, stays at the precision given by the internal points. If the requirement that the scalar product be diagonal is further imposed, then the precision at boundary points drop to half its value.

6.1. **How do we impose the boundary condition at the discrete level?** One would think that one would just take the values of the solution at the boundary as prescribed by the boundary condition and would not use the time derivative of those solution-components for evolution. But then, the energy calculation would not be valid! The energy condition asumes one uses it at all points!

Instead of that we would change the equation on the boundary by adding a term to the time derivative so as to impose the boundary condition in a weak way. The scheme results,

$$\partial_t v_i = a(D_p v)_i + \delta_{0,i} \frac{(sign(a) - 1)|a|}{2h^{00} dx}(g_0(t) - v_0) + \delta_{N-1,i} \frac{(sign(a) + 1)|a|}{2h^{N-1N-1} dx}(g_1(t) - v_{N-1}).$$

This last term is called a **penalty term**, as $dx \to 0$ that term becomes dominant and drives, through the equation, the solution at the boundary to becomes the boundary value. In our example $a > 0$ so we just keep the last term for the next computation.

$$\partial_t v_i = a(D_p v)_i + \delta_{N-1,i} \frac{a}{h^{N-1N-1} dx}(g(t) - v_{N-1})$$

Recalling that in this case,

$$\mathcal{E}_D(t) := \frac{1}{2} \sum_{l=1}^{N} h^{ij} v_i v_j \, dx,$$

we get,

$$\begin{aligned}
\frac{d\mathcal{E}_D(t)}{dt} &= \frac{1}{2} \sum_{l=1}^{N} h^{ij} [a(Dv)_i v_j + v_i a(Dv)_j] \, dx + v_{N-1} a(g(t) - v_{N-1}) \\
&= \frac{a}{2} [v_{N-1} v_{N-1} - u_0 v_0] + v_{N-1} a(g(t) - v_{N-1}) \\
&\leq \frac{a}{2} [v_{N-1}^2 + 2 v_{N-1}(g(t) - v_{N-1})] \\
&\leq \frac{a}{2} [-v_{N-1}^2 + 2 v_{N-1} g(t)] \\
&\leq \frac{a}{2} [-v_{N-1}^2 + v_{N-1}^2 + g(t)^2] \\
&\leq \frac{a}{2} g(t)^2
\end{aligned}$$

(3)

From which an identical estimate as in the continuum case follows.