18.336 Spring 2011
Class notes

Laurent Demanet

Draft February 14, 2011

Preface

In its current form the class 18.336, Computational methods for partial differential equations, covers finite differences, spectral methods, and finite volumes.

Great books are:

- For finite differences, [2].
- For spectral methods, [3].
- For finite volumes, [1].

(List main classes of methods, their pros and cons).

(List main classes of equations, and the main difficulties encountered in solving them.)

Chapter 1

Finite differences, time-independent problems

1.1 Schemes on regular grids

Assume a Cartesian grid $x_j = jh$, with h the grid spacing (also denoted Δx). If a smooth function $u(x)$ is only known through its samples $u_j = u(x_j)$, the simplest approximations to the first derivative $u'(x_j)$ are the following *difference* formulas:

- Forward difference
$$D_+ u_j = \frac{u_{j+1} - u_j}{h}.$$
- Backward difference
$$D_- u_j = \frac{u_j - u_{j-1}}{h}.$$
- Centered difference
$$D_c u_j = \frac{u_{j+1} - u_{j-1}}{2h}.$$

Each of these approximations consists in replacing the slope of the tangent to the graph of $u(x)$ a x, by the slope of some secant supported by neighboring points. *Finite* refers to the fact that the stencil (the set of points where the function evaluation is used) is usually rather compact, at most a few points.

The accuracy of a difference formula depends on how the error scales as a function of h. For instance, if the function u is smooth enough, it holds that

$$|D_+ u_j - u'(x_j)| \leq C\,h,$$

for some number C independent of h (a "constant"). The same inequality holds for the backward difference, possibly with a different constant. As for the centered difference we have

$$|D_c u_j - u'(x_j)| \leq C\,h^2,$$

which is much better when the grid is fine, i.e. as $h \to 0$. It was also clear intuitively that the centered difference would be better, by symmetry.

When the rate of convergence of the error as a function of the grid spacing h scales like h^p, we say that the method has *order of accuracy* p, or *is of order* p. Hence $D_\pm$ is only first-order accurate, while D_c is a second-order method. We also write that the error is $O(h^p)$ (big-O h^p), or in a formula as

$$D_+ u_j - u'(x_j) = O(h).$$

The study of truncation errors is done via Taylor series, and requires that the function $u(x)$ possesses sufficiently many derivatives (one is not enough). For instance, to study the order of accuracy of D_c, write (without loss of generality $x_j = 0$)

$$u(h) = u(0) + hu'(0) + \frac{h^2}{2}u''(0) + \frac{h^3}{6}u'''(0) + \ldots$$

$$u(-h) = u(0) - hu'(0) + \frac{h^2}{2}u''(0) - \frac{h^3}{6}u'''(0) + \ldots$$

subtract the two equations and get

$$\frac{u(h) - u(-h)}{2h} = u'(0) + \frac{h^2}{6}u'''(0) + \ldots$$

To leading order, the error is $\frac{h^2}{6}u'''(0) = O(h^2)$. The proportionality constant involves the third derivative of h, so it is important that the function is sufficiently many times differentiable (here 3 times). You shouldn't worry about the three dots hiding the higher-order terms: you can terminate the Taylor series above via the remainder $\frac{h^3}{6}u'''(\xi)$ for some $\xi \in [0, h]$, so what matters is that the function has a third derivative over the whole subinterval between two grid points.

Watch out: *order of accuracy usually loses its meaning when the function to be differentiated is not smooth.* In practice, the error indeed does not scale as promised by Taylor expansions when the function lacks smoothness. So pushing the order to be large is not the solution to every problem.

For reference, if you need a third-order formula for the first derivative, here it is:

$$D_{+--}u_j = \frac{2u_{j+1} + 3u_j - 6u_{j-1} + u_{j-2}}{6h}.$$

Approximations to the second derivative can be obtained by second differences, notably the three-point rule

$$D_{c,2}u_j = \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2}.$$

It is easy to check that this centered second difference has order of accuracy two. It turns out that $D_{c,2} = D_+D_- = D_-D_+$, but this is not the same thing as D_c^2 (why?).

Difference formulas can be determined by the method of undetermined coefficients: to determine D_{+--} above you would write

$$D_{+--}u_j = \frac{au_{j+1} + bu_j + cu_{j-1} + du_{j-2}}{h},$$

then plug in the Taylor expansions, make sure the first derivative comes out with coefficient 1, and equate as many coefficients of other low powers of h to zero. as possible. This will give rise to a small linear system that needs to be solved. This method is neither fun nor efficient to compute difference formulas!

Another interesting method – which has far-reaching applications to spectral methods – is to build an interpolant and to differentiate the latter. But for compact finite differences schemes it is easiest to look up existing schemes. Most of the useful ones are tabulated.

If you *need* to come up with a new formula yourself, the *calculus of finite difference operators* is an elegant way to do it. We start from the observation that a Taylor formula is in general reminiscent of the Taylor series of the exponential, even when the function is not the exponential. Put $d/dx = D$ for simplicity of notations:

$$\begin{aligned} u(h) &= u(0) + hu'(0) + \frac{h^2}{2}u''(0) + \frac{h^3}{6}u'''(0) + \ldots \\ &= \left[I + hD + \frac{1}{2}(hD)^2 + \frac{1}{3!}(hD)^3 + \ldots \right] u(0) \\ &=: \exp(hD)u(0). \end{aligned}$$

In the last we took the freedom of taking the exponential of an operator, like you would take the exponential of a matrix; the definition of the exponential of an operator is precisely the Taylor series itself. When we write $Au(x)$ for an operator A, we mean to apply A to u, and only then evaluate the answer at x.

BEGIN PARENTHESIS. This notion of operator exponential also comes up when solving time-dependent problems. If there is an operator A such that $u_t = Au$ (where the subscript t denotes time differentiation), then the solution is $u = \exp(tA)u_0$, where u_0 is the initial condition. You may have seen this being done for ODE, where A is a matrix. In the context of PDE, the operator A (the generator) usually contains the boundary conditions. In the special case of $A = D = \frac{\partial}{\partial x}$, we say that the derivative is the generator of translations,

$$u_t(x,t) = \frac{\partial}{\partial x}u(x,t), \qquad u(x,0) = u_0(x)$$

$$\Leftrightarrow$$

$$u(x,t) = \exp(t\frac{\partial}{\partial x})u_0(x) = u_0(x+t).$$

Incidentally, we have just solved a one-way wave equation on the real line. END PARENTHESIS.

So $u(h) = \exp(hD)u(0)$, and the forward difference D_+ (say) itself provides a similar formula. D_+ can be defined to act on the whole function u (and not simply on its samples) as

$$D_+u(x) = \frac{u(x+h) - u(x)}{h}.$$

As a consequence,

$$u(h) = (1 + hD_+)u(0),$$

so we conclude

$$\exp(hD) = 1 + hD_+.$$

A leap of faith leads us to write[1]

$$hD = \log(1 + hD_+).$$

We can now return to a Taylor series, which is actually the definition of the logarithm of an operator. Since h is small, and putting convergence questions aside, it is sensible to write

$$hD = \log(1 + hD_+) = hD_+ - \frac{(hD_+)^2}{2} + \frac{(hD_+)^3}{3} - \frac{(hD_+)^4}{4} + \ldots$$

By truncating the right-hand side at different places, we get different approximations of hD. The resulting order of accuracy is precisely the order of the highest power of h kept in the partial Taylor series. The forward difference formula is recovered to first order. To second order,

$$\begin{aligned} u'(0) &\simeq D_+u(0) - \frac{h}{2}D_+^2u(0) \\ &= \frac{u(h) - u(0)}{h} - \frac{h}{2}\frac{D_+u(h) - D_+u(0)}{h} \\ &= \frac{u(h) - u(0)}{h} + \frac{1}{2}\left(\frac{u(2h) - u(h)}{h} - \frac{u(h) - u(0)}{h}\right) \\ &= \frac{-\frac{1}{2}u(2h) + 2u(h) - \frac{3}{2}u(0)}{h}. \end{aligned}$$

This is a second-order one-sided formula for the first derivative. It is particularly useful at the left boundary of an interval (or a rectangle in 2D, etc.), when the sample $u(-h)$ is not available, but where a first-order formula isn't accurate enough. Low order at the boundaries almost always offsets the advantage of being high order inside!

The calculus of operators can be pushed further. For instance, define the half-width centered difference

$$\delta u(x) = u(x + \frac{h}{2}) - u(x - \frac{h}{2}).$$

[1] This would be justified by inverting the Taylor series, which is sometimes done in advanced calculus classes. Look up the Lagrange-Brmann formula if you're interested.

This operation is not permitted when only the samples at $x+nh$ are available for n integer, but certain combinations of it are useful, like

$$\delta^2 u = h^2 D_{c,2} u.$$

This δ can also be used in place of D_+ in the exp - log argument above. We have

$$\delta u = \exp(hD/2)u - \exp(-hD/2)u = 2\sinh(hD/2)u,$$

so, by inversion,

$$\frac{hD}{2} = \sinh^{-1}\left(\frac{\delta}{2}\right).$$

The Taylor series of the “area hyperbolic sine” can be used to give

$$\frac{hD}{2} = \delta - \frac{\delta^3}{24} + \frac{3\delta^5}{640} - \frac{5\delta^7}{7168} + \dots$$

This is not a useful formula per se, since the odd powers of δ don’t fall on the grid. However, squaring the expression gives

$$(hD)^2 = \delta^2 - \frac{\delta^4}{12} + \frac{\delta^6}{90} - \frac{\delta^8}{560} + \dots$$

which is a suitable expansion of the second derivative. The first term recovers $h^2 D_{c,2}$, and by truncating elsewhere we can obtain higher-order formulas.

Generalizations of the centered difference formula for the first-derivative make a good homework question.

1.2 A one-dimensional boundary value problem

Consider the heat equation, which links a temperature profile $u(x,t)$ to a heat source/sink $f(x,t)$. In the interval $[0,1]$t is written

$$u_t(x,t) = (\kappa(x)u_x)_x + f(x,t), \qquad x \in [0,1],$$

where $\kappa(x)$ is the local heat conductivity, and adequate boundary conditions such as $u(0) = \alpha$, $u(1) = \beta$ need to be enforced. In what follows we’ll assume

$\kappa(x) = 1$. Further assuming that f does not depend on time, the temperature will reach an equilibrium (in the "steady state") to a profile $u(x)$ solution of

$$-u''(x) = f(x), \qquad u(0) = \alpha,\, u(1) = \beta.$$

This is the simplest boundary-value problem, and also the simplest elliptic equation. Of course it has an explicit solution, but the properties of its discretization must be understood before anything else involving finite differences.

On a Cartesian grid $x_j = jh$ with $x_0 = 0$, $x_{n+1} = 1$ and $h = 1/(n+1)$, we have n interior unknowns u_j with $1 \leq j \leq n$. Using the three-point rule, we get

$$\frac{-U_{j+1} + 2U_j - U_{j-1}}{h^2} = f(x_j), \qquad U_0 = \alpha,\, U_{n+1} = \beta.$$

The numbers $U_1, \ldots, U_n$ are therefore the solution u of the linear system $KU = F$, where F is the vector of samples $f(x_j)$, and K is the all-important matrix (when $n = 4$)

$$K = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}.$$

However large n may be, the linear system is best solved by Gaussian elimination. It provides a factorization of K as LM with L lower triangular and M upper triangular. The complexity of this factorization is an optimal $O(n)$, because the factors L and M are both bidiagonal. Gaussian elimination preserves the sparsity pattern of K in the one-dimensional case. Once L and M are available, the systems $Lx = F$ and $MU = x$ are solved by forward substitution and back substitution respectively. We will return to the higher-dimensional case in the sequel.

How good of an approximation to the true solution u is provided by the solution u_j to this system? We are interested in finding a bound on the difference $e_j = U_j - u(x_j)$ which will depend on the choice of grid spacing h. Call $E = U - u$ the vector of errors E_j. In what follows we will need the 2-norm of a vector such as E:

$$\|E\| = \left[h \sum_{j=1}^{n} |E_j|^2 \right]^{1/2}.$$

The place of h is important in this definition: it ensures that $\|E\|$ has the same order of magnitude as its entries, when they are comparable. Furthermore, sequences of values indexed by j only have a life as discretized versions of continuous quantities. With the convention above we recognize a quadrature formula for the continuous L^2 norm.

The maximum (or uniform, ℓ_∞) norm is also useful:

$$\|E\|_\infty = \max_j |E_j|.$$

Two aspects will play a role in quantifying $\|E\|$:

- *Consistency*, or what is the order of the numerical scheme; and
- *Stability*, or how do consistency errors at the level of the equation transfer to actual errors at the level of its solution.

We will see that, together, consistency and stability guarantee *convergence*. This will be a recurring theme in the study of time-dependent problems as well.

1.3 Convergence theory

1.3.1 Consistency

In short, consistency is the idea that the FD scheme works with some order of accuracy for every term in a PDE, including at the boundaries.

The convergence question is to quantify the extent to which *the numerical solution U_j does not provide the right values for the exact equation.* Establishing consistency is a first step in that direction. It answers the simpler question of quantifying the extent to which *the samples $u(x_j)$ do not provide the right values for the numerical scheme.*

Consistency is usually easy to establish if we have smoothness assumptions on the solution. For the example in the previous section, if we know $u \in C^4$, then

$$\frac{-u(x_{j+1}) + 2u(x_j) - u(x_{j-1})}{h^2} - f(x_j) = -u''(x_j) + O(h^2) - f(x_j) = O(h^2),$$

where the coefficient of h^2 involves $u^{(4)}$. Consistency is the property that the $O(h^2)$ remainder above, called *local truncation error*, tends to zero as the mesh is refined ($h \to 0$). The same property should be true of the one-sided schemes at the boundary.

More generally, we deal with linear systems $A^h U^h = F^h$, where the superscript h is used to remind that the vectors and matrices have sizes that depend on the grid spacing h. We then have the following definition.

Definition 1. *Suppose a finite difference scheme for some boundary-value problem gives a sequence of linear system $A^h U^h = F^h$. Denote by τ^h the local truncation error,*

$$\tau^h = A^h u^h - F^h.$$

The scheme is said to be consistent *if $\|\tau^h\| \to 0$ as $h \to 0$.*

1.3.2 Stability

A small local truncation error τ^h does not necessarily imply a small (actual) error E. The relationship between the two quantities is obtained as follows.

$$A^h U^h = F^h,$$

$$A^h u^h = F^h + \tau^h.$$

Subtract to obtain

$$A^h E = -\tau^h.$$

Hence to obtain E from τ^h, it suffices to solve the same discretized BVP as was originally given. This gives

$$E = -(A^h)^{-1}\tau^h.$$

In order to guarantee that E and τ^h be of the same order of magnitude, it suffices to control the corresponding (induced) matrix norm of $(A^h)^{-1}$. The ℓ_2 matrix norm of a matrix M is defined as

$$\|M\| = \max_{x \neq 0} \frac{\|Mx\|}{\|x\|}.$$

This norm is used in the following definition.

Definition 2. *Suppose a finite difference scheme for some boundary-value problem gives a sequence of linear system* $A^h U^h = F^h$*. The scheme is said to be stable if there exist* $h^* > 0$ *and* $C > 0$ *such that, for all* $0 < h < h^*$,

$$\|(A^h)^{-1}\| \leq C.$$

So we see that a scheme is convergent ($\|E\| \to 0$) if we can show that it is both consistent and stable. (Note that the definition above is sufficient but not necessary.) That convergence follows from consistency and stability is a recurring theme in FD discretizations.

To control $\|(A^h)^{-1}\|_2$, it is useful to have information about the eigenvalues of A^h. If M is a symmetric matrix, it holds that

$$\|M\| = \max_j |\lambda_j(M)| = \rho(M),$$

which is called the spectral radius. The inverse M^{-1} is also symmetric, so

$$\|M^{-1}\| = \frac{1}{\min_j |\lambda_j(M)|}.$$

If M is not symmetric, the spectral radius is in general smaller than the norm, but equality can be restored if we consider singular values:

$$\|M\| = \max_j |\sigma_j(M)|,$$

where $\sigma_j(M) = \sqrt{\lambda_j(M^T M)}$. Analogously, for square invertible matrices we also have

$$\|M^{-1}\| = \frac{1}{\min_j |\sigma_j(M)|}.$$

1.4 The spectrum of the discrete Laplacian

Let us discuss the (very important) spectrum of $-d^2/dx^2$ with Dirichlet boundary conditions, in order to establish stability of the centered FD scheme for our simple BVP.

The matrix of interest is K/h^2, where K was defined earlier. It is a symmetric matrix, so it suffices to find the eigenvalues. Let us get a good intuition of what the eigenvectors look like by returning to the continuous eigen-problem

$$-v'' = \lambda v, \qquad v(0) = v(1) = 0.$$

The eigenfunctions are

$$v_m(x) = \sin(m\pi x),$$

for $m \geq 1$, with eigenvalue

$$\lambda_m = \pi^2 m^2.$$

It is perhaps a stroke of luck that the eigenvectors of the n-by-n matrix K/h^2 are precisely the discretized sines on the Cartesian grid $x_j = jm$,

$$v_j^m = \sin(m\pi x_j), \qquad 1 \leq j, m \leq n.$$

The application of two trigonometric formulas reveals that the eigenvalues are given by

$$\lambda_m = \frac{4}{h^2} \sin^2\left(\frac{m\pi h}{2}\right), \qquad 1 \leq m \leq n. \tag{1.1}$$

Two regimes must be contrasted.

- *Low frequencies.* For small m, and small h, a Taylor expansion of (1.1) reveals

  $$\lambda_m = \pi^2 m^2 + O(h^2), \qquad m \geq 1, \text{ small}.$$

  This matches the continuous eigenvalues to second order. The match is intuitively good because, for small m, the eigenfunctions $\sin(m\pi x)$ are properly resolved on the grid, and the computation of the second derivative by second difference is accurate.

- *High frequencies.* The match of eigenvalues worsens as m gets large. The largest m is equal to n, for which

  $$\lambda_n = \frac{4}{h^2} + O(h^2).$$

  Regardless of h, the modes for which m is close to n correspond to eigenfunctions that are not properly sampled, in the sense that only a few grid points are used per wavelength (somewhere between 2 and 4, say.) At two points per wavelength (ppw), a mode is called flip-flop. For some applications such a coarse sampling could be OK – after all the Shannon sampling theorem guarantees the possibility of spectral interpolation all the way down to 2 ppw – but numerical differentiation via the second difference is definitely not accurate below about 4 ppw. As a result the discrete eigenvalues are not faithfully representative of the continuous case.

In short, the 3-point FD for $-d^2/dx^2$ with Dirichlet boundary conditions is positive definite, with smallest eigenvalue $\simeq \pi^2$, and largest eigenvalue $\simeq 4/h^2$. The stability criterion defined in the previous section is obeyed in this case, because the smallest eigenvalue does not tend to zero as $h \to 0$.

We will return to the important spectral properties of $-d^2/dx^2$ in the scope of convergence speed for iterative solvers, as well as numerical dispersion relations for time-dependent wave equations.

1.5 Linear solvers

Let us now address the problem of solving the system $AU = F$ numerically.

1.6 Preconditioning, multigrid

Chapter 2

Finite differences, time-dependent problems

2.1 Stability analysis

2.2 Numerical dispersion

Chapter 3

Spectral methods

3.1 Periodic grids

Interpolation, differentiation Poisson summation formula

3.2 Chebyshev grids

Chapter 4

Finite volumes

4.1 Conservation laws and shocks

4.2 Godunov, Riemann problems, limiters, TVD methods, etc.

Bibliography

[1] R. LeVeque, *Finite volume methods for hyperbolic problems*, Cambridge University Press, 2002.

[2] R. LeVeque, *Finite difference methods for ordinary and partial differential equations*, SIAM, 2007.

[3] L. N .Trefethen, *Spectral methods in Matlab*, SIAM, 2000.