B6.2 Numerical Solution of Differential Equations II (2018-2019)

Lecture Notes

Dr Ricardo Ruiz Baier

1. Lecture 1

We consider second order ordinary differential equations (ODEs) involving boundary conditions, given by

$$\frac{d^2y}{dx^2} = y'' = f(x, y, y') \quad \text{with boundary conditions} \quad y(a) = \alpha,\ y(b) = \beta \tag{1}$$

and seek a solution $y(x)$ for $x \in [a, b]$. Boundary value ODEs have a more sophisticated existence and uniqueness theory as compared to initial value problems. We omit this literature and instead focus on methods for the approximate solution to boundary value ODEs when the cases where solutions exist and are unique.

In this course we consider two conceptually different approaches to construct approximate solutions within any prescribed accuracy. The first approach transforms the boundary value problem into initial value problem(s), allowing approximate solutions to be computed using methods such as from the class of Runge Kutta and linear-multistep methods; this approach is broadly termed "shooting methods" and will be the focus of this lecture. The second approach involves explicit discretisation of the x variable, approximating the difference operators by matrices, and solving the resulting system of equations. This second approach is more typical of methods used throughout this course for boundary value partial differential equations (PDEs).

1.1. Shooting method for linear ODEs.

Before considering a numerical method for computing approximate solutions to ODEs we illustrate the principle of the shooting method for linear second order ODE boundary value problems (BVPs) of the form

$$y'' = p(x)y' + q(x)y + r(x) \quad \text{with boundary conditions} \quad y(a) = \alpha,\ y(b) = \beta \tag{2}$$

for $x \in [a, b]$. From this boundary value problem we construct two initial value problems using the same coefficient functions $p(x)$, $q(x)$, and $r(x)$:

$$y'' = p(x)y' + q(x)y + r(x) \quad \text{with b. c.} \quad y(a) = \alpha,\ y'(a) = 0 \tag{3}$$

and

$$y'' = p(x)y' + q(x)y \quad \text{with b. c.} \quad y(a) = 0,\ y'(a) = 1. \tag{4}$$

These two IVPs can be solved within arbitrary precision using any of the standard numerical techniques, such as Runge Kutta methods. From these two IVP solutions it is possible to construct an approximate solution of the BVP (2) by taking a linear combination. Let $y_1(x)$ be an approximate solution to (3), let $y_2(x)$ be an approximate solution to (4), and set $y(x) = y_1(x) + \gamma y_2(x)$. By construction $y(a) = \alpha$ as required. To satisfy the second boundary value one needs $y(b) = \beta = y_1(b) + \gamma y_2(b)$, which can be satisfied by selecting $\gamma = (\beta - y_1(b))/y_2(b)$. This approach is effective provided $y_2(b)$ is well separated from zero, allowing the BVP to be solved approximately by instead solving two related IVPs.

1.2. Shooting method for nonlinear ODEs.

Nonlinear BVPs cannot typically be transformed into a pair of linear IVPs. However, a similar approach can be devised. Rather than solving (1), one can replace the right boundary condition with a user specified slope at the left boundary

$$\frac{d^2y}{dx^2} = y'' = f(x, y, y') \quad \text{with b. c.} \quad y(a) = \alpha,\ y'(a) = s, \tag{5}$$

giving a parametrised solution, $y(x; s)$, for each s. It then remains to find a value of s, say s^*, such that its parametrisation matches the right boundary condition $y(b; s^*) = \beta$ within the specified accuracy. For BVPs with unique solutions, the IVP satisfying $y(b; s^*) = \beta$ necessarily has the same solution as the BVP we seek to approximately solve.

Solving (1) has been reduced to solving for the s which solves $\phi(s) := y(b;s) - \beta = 0$, a standard root finding problem. This root finding problem is particularly tractable for IVPs that can be well approximated numerically; specifically, $\phi(s)$ must be a continuous function. Methods well suited for computing the root of $\phi(s)$ may include: the bisection method (provided $\phi(s)$ changes sign) and the Secant method. These root finding methods simply require function evaluation of $\phi(s)$, which can be well approximated using standard numerical methods for IVPs; however, it is worth noting that though evaluating $\phi(s)$ is straightforward, it may be computationally intensive. For an overall computationally efficient solution to (1) we need a root finding method that requires few iterates. Newton's method is particularly efficient, quadratically convergent, when an initial estimate to the root is available.

Newton's method for $\phi(s) := y(b;s) - \beta = 0$ is given by

$$s^{n+1} = s^n - \frac{y(b;s^n) - \beta}{y_s(b;s^n)} \tag{6}$$

where $y_s(b;s^n)$ is the derivative of $y(b;s)$ with respect to s, evaluated at s^n. The function $y_s(b;s)$ is not readily available, but can be approximated as follows. Applying $\frac{\partial}{\partial s}$ to the ODE in (5) gives

$$y''_s = f_x x_s + f_y y_s + f_{y'} y'_s.$$

Noting that $x_s = 0$ due to x being independent of s, applying $\frac{\partial}{\partial s}$ to the initial conditions in (5), and setting $z(x;s) = y_s(x,s)$ for ease of notation gives an additional second order IVP

$$\begin{aligned} \frac{d^2 z}{dx^2} = \quad z'' \quad &= f_y(x, y(x;s), y(x;s)')z + f_{y'}(x, y(x;s), y'(x;s))z' \\ &\text{with boundary conditions } \; z(a) = 0, \; z'(a) = 1. \end{aligned} \tag{7}$$

It is important to note that the coefficients $f_y(x,y,y')$ and $f_{y'}(x,y,y')$ require both the user to be able to compute these derivatives of $f(x,y,y')$, and require an approximate solution of $y(x;s)$ and $y'(x;s)$ for each value of x used in computing the approximate solution to (7).

2. Lecture 2

2.1. **Finite difference method for second order linear ODEs.** We express the (2) linear differential equation by

$$L(y) = -y'' + p(x)y' + q(x)y = -r(x) \quad \text{with b. c.} \quad y(a) = \alpha,\ y(b) = \beta \tag{8}$$

for $x \in [a, b]$. The finite difference method begins by discretizing x using an equally spaced grid

$$x_j = a + jh \quad \text{with} \quad h = \frac{b-a}{n+1}, \quad \text{for } j = 0, 1, \ldots, n+1.$$

Let y_j be our approximation to $y(x_j)$, we can approximate the differential operator $L(y)$ with suitable finite difference approximations to the derivatives. For a three point stencil (using just three points per equation) we approximate

$$y''(x_j) = \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} - \frac{1}{12}h^2 y^{(4)}(\xi_j)$$

and

$$y'(x_j) = \frac{y_{j+1} - y_{j-1}}{2h} - \frac{1}{6}h^2 y^{(3)}(\eta_j).$$

The resulting approximation to (8) at x_j is (after multiplication by $\frac{1}{2}h^2$)

$$L_h(y_j) = a_j y_{j-1} + b_j y_j + c_j y_{j-1} = -r(x_j) \quad \text{for j=1,2,...n} \tag{9}$$

where

$$\begin{aligned} a_j &:= -\frac{1}{2}\left[1 + \frac{1}{2}hp(x_j)\right] \\ b_j &:= \left[1 + \frac{1}{2}h^2 q(x_j)\right] \\ c_j &:= -\frac{1}{2}\left[1 - \frac{1}{2}hp(x_j)\right] \end{aligned} \tag{10}$$

and boundary conditions $y_0 = \alpha$ and $y_{n+1} = \beta$. The n unknowns, y_j for $j = 1, 2, \ldots, n$, can then be cast as a linear system of equation

$$Ay = -r - a_1 \alpha e_1 - c_n \beta e_n \tag{11}$$

where: e_ℓ is the unit n vector with value $e_\ell(k) = 1$ if $\ell = k$ and zero otherwise, r is the vector with entries $\frac{1}{2}h^2 r(x_j)$, A is the $n \times n$ tridiagonal matrix with values b_j on the diagonal for $j = 1, 2, \ldots, n$, a_j on the sub-diagonal for $j = 2, 3, \ldots, n$, and c_j on the super-diagonal for $j = 1, 2, \ldots, n-1$, and y the vector with entries y_j.

Our numerical method for solving for an approximate solution to (8) (on the grid x_j) is now cast as the solution of a linear system. The central questions to resolve for this method are:

- Does the linear system (11) have a unique solution?
- What is the computational cost of solving the system (11)?
- At what rate does the error $\max_j |y(x_j) - y_j|$ converge to zero as h decreases to zero? (This is referred to as the order of accuracy.)

To address invertibility we impose conditions on the ODE variable coefficient function $q(x)$ to have

$$\min_{x \in [a,b]} q(x) = Q_* > 0 \tag{12}$$

and that the stepsize is sufficiently small compared to the maximum of the coefficient function $p(x)$

$$h < \frac{2}{P^*} \quad \text{where} \quad P^* = \max_{x \in [a,b]} p(x). \tag{13}$$

The first condition ensure that the diagonal values in A are greater than one, $b_j \geq 1 + \frac{1}{2}h^2 Q_*$. The second condition ensures that he sum of the off diagonal entries in A have magnitude 1, $|a_j| + |c_j| = 1$. Gershgorin disc theorem using these two facts tell us that the n eigenvalues of A are contained in discs of radius 1 centred at b_j. As b_j are greater than one, the discs do not include the origin, ensuring that zero is not an eigenvalue of A. Moreover, A is diagonally dominant, and can be easily solved using Gaussian Elimination without need for pivoting. This later fact tells us that a stable solution can be computed in order n operations. We have now verified that, with the conditions imposed, the linear system corresponding to our method to solve an approximate solution to (8) has a unique solution and can be solved efficiently.

It then remains to establish the order of accuracy for our method. We begin by noting the truncation error for L_h that results from the finite difference approximations to the differential operator; simple Taylor series expansions show

$$L_h(y(x_j)) - L(y(x_j)) = \frac{-h^2}{12}\left[y^{(4)}(\xi_j) - 2p(x_j)y^{(3)}(\eta_j)\right]. \tag{14}$$

This shows that on the mesh x_j, the solution to the ODE, $y(x_j)$ gives the same answer to differential operator L and the finite difference operator L_h to withing $\mathcal{O}(h^2)$. In order to establish that y_j is close to $y(x_j)$ we also need to ensure that the finite difference operator L_h is "stable." We refer to a finite difference operator L_h as stable with factor M if there exists a finite M such that

$$\max_j |\nu_j| \leq M\left\{\max\left(|\nu_0|, |\nu_{n+1}|\right) + \max_j |L_h\nu_j|\right\}. \tag{15}$$

Noting that

$$\begin{aligned} L_h y_j - L_h y(x_j) &= -r(x_j) - L_h y(x_j) \\ &= Ly(x_j) - L_h y(x_j) \end{aligned} \tag{16}$$

and using the truncation error bound (14) gives the bound

$$|L_h(y_j - y(x_j))| = |Ly(x_j) - L_h y(x_j)| \leq \frac{h^2}{12}\left|y^{(4)}(\xi_j) - 2p(x_j)y^{(3)}(\eta_j)\right|. \tag{17}$$

Consequently, if L_h is M stable then using $\nu_j = y_j - y(x_j)$ in (15) gives

$$\max_j |y_j - y(x_j)| \leq \frac{Mh^2}{12}\left[\max_{x\in[a,b]} |y^{(4)}(x)| + 2P^* \max_{x\in[a,b]} |y^{(3)}(x)|\right],$$

proving second order approximation rate for the method.

It then remains to show that L_h is a stable operator. To prove this we recall that the operator satisfies

$$b_j y_j = -a_j y_{j-1} - c_j y_{j+1} + \frac{1}{2}h^2 L_h y_j$$

The right hand side can be bounded from above by using the triangle inequality, noting that under the conditions (12) and (13), that $|a_j| + |c_j| = 1$, so taking the max over j on the right hand side gives the upper bound

$$|b_j y_j| \leq \max_j |y_j| + \frac{1}{2}h^2 \max_j |L_h y_j|.$$

The left hand side can be bounded below by $(1 + \frac{1}{2}h^2 Q_*)||y_j|$ for each j, and consequently is also true for the j where the max of $|y_j|$ is achieved. The resulting bound

$$(1 + \frac{1}{2}h^2 Q_*) \max_j |y_j| \leq \max_j |y_j| + \frac{1}{2}h^2 \max_j |L_h y_j|,$$

can be rearranged to

$$\max_j |y_j| \leq \frac{1}{Q_*} \max_j |L_h y_j|,$$

and hence L_h is stable with factor $M = Q_*^{-1}$. Combined with our prior analysis we have proven that the solution to our finite difference approximation is a second order accurate approximation to the true solution.

3. Lecture 3

In this lecture we consider finite difference methods for nonlinear BVPs.

3.1. **Finite difference methods for nonlinear BVPs.** We return to nonlinear second order BVPs (5), here written as

$$L(y) = -y'' + f(x, y, y') = 0 \quad \text{with b. c.} \quad y(a) = \alpha,\ y(b) = \beta. \tag{18}$$

Nonlinear Truncation Error Let us derive a finite difference method for its approximate solution. We begin by replacing the differential operators with finite difference approximations, here keeping to a three point stencil.

$$L_h(y_j) = -\frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} + f\left(x_j, y_j, \frac{y_{j+1} - y_{j-1}}{2h}\right) \quad \text{for} \quad j = 1, 2, \ldots, n \tag{19}$$

with boundary values $y_0 = \alpha$ and $y_{n+1} = \beta$. The finite difference operator acting on the approximate solution y_j is within $\mathcal{O}(h^2)$ of the finite difference operator acting on the true solution on the corresponding mesh, $y(x_j)$. This truncation error is given by:

$$\begin{aligned} L_h y(x_j) - L_h y_j &= L_h y(x_j) - L y(x_j) \\ &= -\frac{y(x_{j+1}) - 2y(x_j) + y(x_{j-1})}{h^2} + y''(x_j) \\ &+ f\left(x_j, y(x_j), \frac{y(x_{j+1}) - y(x_{j-1})}{2h}\right) - f\left(x_j, y(x_j), y'(x_j)\right) \\ &= \frac{-1}{12} h^2 y^{(4)}(\xi_j) + \frac{1}{6} h^2 f_{y'}(x_j, y(x_j), y'(x_j)) y^{(3)}(\eta_j) \\ &= \frac{h^2}{12}\left[-y^{(4)}(\xi_j) + 2 f_{y'}(x_j, y(x_j), y'(x_j)) y^{(3)}(\eta_j)\right] \end{aligned} \tag{20}$$

where the $f_{y'}$ notation indicates partial derivative of f with respect to its third argument, and the equality is determined by using previous differences of the differential and difference operators. It now remains to show that a) the operator is *stable* so that $\max_j |L_h y(x_j) - L_h y_j|$ being proportional to $\mathcal{O}(h^2)$ implies that $\max_j |y(x_j) - y_j|$ is similarly second order in h^2, and b) to show that the finite difference system (19) has a solution, which we are able to find using standard root finding techniques.

Nonlinear Stability We have shown a second order truncation error (20) for the finite difference scheme (19). In order to show that $\max_j |y_j - y(x_j)|$ is of the same order as the truncation error we repeat a stability analysis of the finite difference operator $L_h(\cdot)$. When considering linear operators the notion of stability was given in terms of a single vector (15); here the non-linearity of the operator requires a slightly more general definition of stability, given in terms of two vectors. We refer to a finite difference operator L_h as stable with factor M if there exists a finite M such that

$$\max_j |u_j - v_j| \le M \left\{ \max\left(|u_0 - v_0|, |u_{n+1} - v_{n+1}|\right) + \max_j |L_h u_j - L_h v_j| \right\}. \tag{21}$$

For linear operators L_h, the definition (21) recovers the prior definition (15).

We first establish that if L_h is stable, then the error is bounded by the truncation error. If L_h is stable with factor M then

$$\begin{aligned} \max_j |y_j - y(x_j)| &\le M \max_j |L_h y_j - L_h y(x_j)| \\ &= M \max_j |L y(x_j) - L_h y(x_j)| \end{aligned} \tag{22}$$

where the last equality uses that $L_h y_j$ is defined to be equal to $Ly(x_j)$. The right hand side of (22) is simply M times the truncation error for the finite difference operator, which for (19) we have shown to be second order, $\mathcal{O}(h^2)$. It then remains to show that L_h is stable, under suitable conditions on $f(\cdot,\cdot,\cdot)$.

In order to show stability of (19) we use vector Taylor series:

$$\begin{aligned}
L_h u_j - L_h v_j &= -\frac{u_{j+1}-2u_j+u_{j-1}}{h^2} + \frac{v_{j+1}-2v_j+v_{j-1}}{h^2} \\
&\quad + f\left(x_j, u_j, \frac{u_{j+1}-u_{j-1}}{2h}\right) - f\left(x_j, v_j, \frac{v_{j+1}-v_{j-1}}{2h}\right) \\
&= -h^{-2}(u_{j+1}-v_{j+1}) + 2h^{-2}(u_j - v_j) - h^{-2}(u_{j-1}-v_{j-1}) \\
&\quad + \nabla f\left(x_j, u_j + \theta(v_j-u_j), \frac{u_{j+1}-u_{j-1}}{2h}\right. \\
&\qquad \left. + \theta\left[\frac{v_{j+1}-v_{j-1}}{2h} - \frac{u_{j+1}-u_{j-1}}{2h}\right]\right) \\
&\quad \cdot \left(0, u_j - v_j, \frac{u_{j+1}-u_{j-1}}{2h} - \frac{v_{j+1}-v_{j-1}}{2h}\right) \\
&= -h^{-2}(u_{j+1}-v_{j+1}) + 2h^2(u_j - v_j) - h^{-2}(u_{j-1}-v_{j-1}) \\
&\quad f_y(x_j,\xi_j,\eta_j)(u_j - v_j) \\
&\quad + f_z(x_j,\xi_j,\eta_j)(2h)^{-1}(u_{j+1}-v_{j+1}-u_{j-1}+v_{j-1}) \\
&= a_j(u_{j-1}-v_{j-1}) + b_j(u_j - v_j) + c_j(u_{j+1}-v_{j+1}) \qquad (23)
\end{aligned}$$

where ξ_j and η_j are for some $\theta \in (0,1)$ and

$$\begin{aligned}
a_j &= -h^{-2} - (2h)^{-1} f_z(x_j,\xi_j,\eta_j) \\
b_j &= 2h^{-2} + f_y(x_j,\xi_j,\eta_j) \\
c_j &= -h^{-2} + (2h)^{-1} f_z(x_j,\xi_j,\eta_j). \qquad (24)
\end{aligned}$$

To bound $|L_h u_j - L_h v_j|$ we first rearrange the final equality in (23) to

$$b_j(u_j - v_j) = -a_j(u_{j-1}-v_{j-1}) - c_j(u_{j+1}-v_{j+1}) + L_h u_j - L_h v_j. \qquad (25)$$

Before computing the desired bound we impose two conditions on the differential equation, similar to those used in the stability analysis of (9). Imposing that $\max|f_z| \le P^*$ and $h \le \frac{2}{P^*}$ gives $|a_j| + |c_j| = 2h^{-2}$ and imposing that $\min f_y > Q_* > 0$ gives $b_j > 2h^{-2} + Q_*$. Taking absolute values of (25), apply the triangle inequality, and maximize over j gives

$$(2h^{-2} + Q_*) \max_j |u_j - v_j| \le 2h^{-2} \max_j |u_j - v_j| + \max_j |L_h u_j - L_h v_j| \qquad (26)$$

which can be simplified to

$$\max_j |u_j - v_j| \le Q_*^{-1} \max_j |L_h u_j - L_h v_j|$$

which is our desired stability bound with factor Q_*^{-1}. Having established the stability factor and previously the second order truncation error proves that the error, $\max_j |y_j - y(x_j)|$ for the finite difference approximation (19) is of order h^2.

4. Lecture 4

In this lecture we continue our analysis of finite difference methods for nonlinear BVPs, showing that the nonlinear system has a unique solution, and proving a method to solve for the solution without knowledge of a good initial guess.

4.1. **Iterative method for solution of nonlinear systems.** At this stage we have a finite difference method (19) whose solution we have shown is within order h^2 of the solution to the nonlinear differential equation (18). However, we have not shown that a) the nonlinear system (19) does in fact have a solution, and b) if it does have a solution we have not given a method by which we can find (approximately) its solution. We address both of these issues simultaneously by considering the iterative algorithm

$$y_j^{m+1} = (1+w)^{-1}\left[\frac{1}{2}(y_{j-1}^m + y_{j+1}^m) + wy_j^m - \frac{h^2}{2}f\left(x_j, y_j, \frac{y_{j+1}^m - y_{j-1}^m}{2h}\right)\right] \tag{27}$$

where the superscript is an iteration counter, not a power. This iteration is arrived at by solving (19) for y_j from the second order differential operator approximation, then adding wy_j (for some $w \neq 1$) to both sides of the equation, dividing by $(1+w)$, and adding iteration counters of one degree less on the right hand side than on the left hand side. We can further condense this iteration as

$$y^{m+1} = g(y^m) \tag{28}$$

where y^m is the vector with entries y_j^m for $j = 0, 1, \ldots, n+1$. We now wish to show a few properties of the iterations y^m: first that they converge and second that they converge to something that satisfies $y = g(y)$ which necessarily implies that the limit is a solution to the finite difference method (19). In order to show these we need to establish that $g(\cdot)$ is a contraction; that is

$$\|g(u) - g(v)\|_\infty \leq \lambda \|u - v\|_\infty$$

for some $0 \leq \lambda < 1$. This analysis is similar to the stability analysis for (19). Letting $g(u)_j$ denote the j^{th} entry of $g(u)$, then

$$\begin{aligned} g(u)_j - g(v)_j &= (1+w)^{-1}\left[\frac{1}{2}\left((u_{j-1} - v_{j-1}) + (u_{j+1} - v_{j+1})\right) + w(u_j - v_j)\right. \\ &- \left.\frac{h^2}{2}\left(f\left(x_j, u_j, \frac{u_{j+1} - u_{j-1}}{2h}\right) + f\left(x_j, v_j, \frac{v_{j+1} - v_{j-1}}{2h}\right)\right)\right] \\ &= -(1+w)^{-1}\frac{h^2}{2}\left[a_j(u_{j-1} - v_{j-1}) + c_j(u_{j+1} - v_{j+1})\right. \\ &\qquad \left. + (b_j - 2h^{-2}(1+w))(u_j - v_j)\right] \end{aligned} \tag{29}$$

with a_j, b_j, and c_j defined as in (24), though with some other ξ_j and η_j. As in the stability analysis we impose that $\max|f_z| \leq P^*$ and $h \leq \frac{2}{P^*}$ gives $|a_j| + |c_j| = 2h^{-2}$ and (using a bound from above instead) impose that $Q_* \leq \min f_y \leq Q^*$ gives $2h^{-2} + Q_* \leq b_j \leq 2h^{-2} + Q^*$. Moreover, we impose that $w \geq \frac{1}{2}h^2Q^*$ so that $|b_j - 2h^{-2}(1+w)| = 2h^{-2}(1+w) - b_j \geq 0$. Then, applying the triangle inequality to the last equality in (29), and taking the max over j we obtain

$$\|g(u) - g(v)\|_\infty \leq \left(1 - \frac{\frac{1}{2}h^2Q_*}{1+w}\right)\|u - v\|_\infty$$

which proves that $g(\cdot)$ is a contraction with factor

$$\lambda(w) := \left(1 - \frac{\frac{1}{2}h^2Q_*}{1+w}\right) < 1$$

for $w \geq \frac{1}{2}h^2Q^*$. Unfortunately $\lambda(w) = 1 - \mathcal{O}(h^2)$ causing the contraction to occur impractically slow for h small. Even so, this is enough to establish the conditions we sought.

Using $y^{m+1} = g(y^m)$ and the contraction principle it is easy to show that $\|y^{k+1} - y^k\|_\infty \leq \lambda^k \|y^1 - y^0\|_\infty$ and consequently that the sequence is a Cauchy sequence. This implies convergence to a limit point that satisfies $y = g(y)$. Moreover, the limit point must be unique by the counter examples that if y and $\tilde{y}$ are solutions that $|y - \tilde{y}| = |g(y) - g(\tilde{y})| = \lambda|y - \tilde{y}|$ for $\lambda < 1$, which is a contradiction, hence proving that the limit is unique. Lastly, the error satisfy $|y^m - y| = |g(y^{m-1}) - g(y)| \leq \lambda|y^{m-1} - y|$, giving a linear convergence rate, though with the factor λ which is close to one. Although this iteration is impractically slow, it has the advantage that convergence is guaranteed to within arbitrary precision for any starting guess.

Remark. In this lecture we have also defined a finite volume method for a linear boundary value problem in 1D with homogeneous boundary conditions. Lecture notes are not yet available.

5. Lecture 5

In this lecture we consider the Poisson Equation, a linear boundary value PDE. Proof of convergence for our approximation involves a refined version of the previous stability analysis, with this approach more adaptable to complex domains.

5.1. **Poisson Equation.** We define the Poisson Equation as

$$L(u) = u_{xx} + u_{yy} = f(x,y) \quad \text{for} \quad (x,y) \in \Omega \tag{30}$$

and, for the moment, with Dirichlet boundary conditions $u(x,y)$ given for $(x,y) \in \delta\Omega$ where $\delta\Omega$ denotes the boundary of Ω. We consider a finite difference approximation of $L(u)$ using a three point centered difference approximation of the second derivative in both x and y, resulting in a five point stencil,

$$L_h u_{j,k} = \frac{u_{j-1,k} + u_{j+1,k} - 4u_{j,k} + u_{j,k-1} + u_{j,k+1}}{h^2} = f(x_j, y_k) \tag{31}$$

for $(x_j, y_k) \in \Omega/\partial\Omega$ where (x_j, y_k) is a grid with $x_{j+1} - x_j = y_{k+1} - y_k = h$ for all j, k. (For instance, if $\Omega = [a,b]^2$ we can use $x_j = a + jh$ and $y_k = a + kh$ for $h = 1/(n+1)$ and $j, k = 0, 1, \ldots, n+1$; however, we are primarily interested in being able to compute approximate solutions on more complex domains.)

Taylor series, as before, is sufficient to show that

$$\begin{aligned} \tau_{j,k} &= L_h u(x_j, y_k) - L_h u_{j,k} = (L_h - L)u(x_j, y_k) \\ &= \frac{1}{12} h^2 \left(u_{xxxx}(\xi_j, y_k) + u_{yyyy}(x_j, \eta_k) \right). \end{aligned} \tag{32}$$

The equations (31) can be expressed as a linear system $Au = f$ where rows of A have diagonal entries $-4h^{-2}$, the super and sub diagonal entries are typically h^{-2} and depending on interactions with boundary conditions a row will may have up to two additional nonzero entries with values h^{-2}. For (j,k) which correspond to a five point stencil that interacts with the boundary, we use the boundary conditions and adjust the entries in f accordingly; otherwise the entries in f are simply given by $f(x_j, y_k)$. Typically the grid (j,k) is ordered using a Lexicographical ordering, ordering $(j,k) > (p,q)$ if $j > p$ or if $j = p$ and $q > k$. The resulting matrix A has only a small fraction of its entries which are not zero, making it computationally efficient to compute matrix vector products Az for some z. In later lectures we will use this property to design efficient methods for computing approximate solutions to $Au = f$. Invertibility of A will be addressed in a later lecture.

We now introduce the *maximum principle*, a technique to show that $\max_{j,k} |u(x_j, y_k) - u_{j,k}| \leq Const.\tau_{max}$ where

$$\tau_{max} = \max_{j,k} |\tau_{j,k}|$$

and $Const.$ is independent of h. The maximum principle uses a *comparison function* $\Phi(x,y)$ designed to allow us to analyse the error

$$e_{j,k} = u(x_j, y_k) - u_{j,k}.$$

We will design the comparison function to have the properties that $L\Phi(x,y) = L_h\Phi(x_j, y_j) = C$ a constant, and that $\Phi(x,y) \geq 0$. We then add a multiple of $\Phi(x,y)$ to the error

$$\psi_{j,k} = e_{j,k} + \alpha\Phi(x_j, y_k)$$

for $\alpha > 0$. Applying the finite difference operator L_h to $\psi_{j,k}$ gives

$$L_h\psi_{j,k} = Le_{j,k} + \alpha L\Phi(x_j, y_k) = \tau_{j,k} + \alpha C.$$

If we select $\alpha = C^{-1}\tau_{max}$ we have that $L_h\psi_{j,k} = \tau_{j,k} + \tau_{max} \geq 0$. As L_h is taking the difference of $\psi_{j,k}$ and the average its four neighbours, $L_h\psi_{j,k} \geq 0$ implies that $\psi_{j,k}$ cannot exceed the max of

the four neighbours used in $L_h\psi_{j,k}$. This property is true for each j,k in which $(x_j,y_k)\in\Omega/\partial\Omega$. Consequently the max of $\psi_{j,k}$ must occur at a boundary point

$$\max_{j,k}\psi_{j,k}\leq\max_{(x_j,y_k)\in\partial\Omega}\psi_{j,k}.$$

For Dirichlet boundary conditions $e_{j,k}$ is zero on the boundary, so $\max_{j,k}\psi_{j,k}\leq\max_{j,k}\Phi(x_j,y_k)=\Phi^*$, where the last equality is our definition of the max of $\Phi(\cdot,\cdot)$. Moreover, as $\Phi(x_j,y_k)\geq 0$, we have that

$$\max_{j,k}e_{j,k}\leq\max_{j,k}\psi_{j,k}=\alpha\Phi^*=C^{-1}\Phi^*\tau_{max}.$$

An example comparison function suitable for this example is $\Phi(x,y)=(x-x_c)^2+(y-y_c)^2$ where (x_c,y_c) is a point such that $\max_{(x,y)\in\Omega}\Phi(x,y)$ is minimized; for this comparison function $L\Phi(x,y)=4$ and $\Phi^*=(a^2+b^2)/4$ where $\Omega\subset[x_c-a/2,x_c+a/2]\times[y_c-b/2,y_c+b/2]$. Implementing these bounds gives

$$\max_{j,k}e_{j,k}\leq\frac{a^2+b^2}{16}\tau_{max}.$$

Repeating the above for $\phi_{j,k}=-e_{j,k}+\Phi(x_j,y_k)$ establishes that

$$\min_{j,k}e_{j,k}=\max_{j,k}-e_{j,k}\leq\max_{j,k}\psi_{j,k}\leq C^{-1}\Phi^*\tau_{max}$$

which when combined with our prior bound gives the desired bound on the error

$$\max_{j,k}|e_{j,k}|\leq\frac{a^2+b^2}{16}\tau_{max}=\mathcal{O}(h^2).$$

6. Lecture 6

In this lecture we return to the question of invertibility of the matrix associated with the system of equations (31). We will also consider alternative finite difference approximations and the impact of domains that do not align perfectly with a regular equispaced grid.

6.1. **Poisson Equation: invertibility.** For rectangular domains Ω is is straightforward to repeat the eigen-analysis of the matrix associated with the system (31) and to show that the eigenvalues are bounded away from zero. Unfortunately this approach does not extend well to more general domains where the eigen-functions of the Laplacian are typically unknown. Here we show that the resulting matrix is invertible by employing a refined version of Gershgorin's Disc Theorem.

Definition 6.1. *An $m \times m$ matrix A is referred to as reducible if there exist sets I and J with the properties that $I \bigcup J = 1, 2, \ldots, m$, and $I \bigcap J = \emptyset$, with $a_{ij} = 0$ for all $i \in I$ and $j \in J$. If A is not reducible we refer to it as irreducible. Moreover, A is referred to as irreducible diagonally dominant (IRDD) if it is weakly row diagonally dominant with at least one row being strictly diagonally dominant.*

Lemma 6.1. *If A is irreducible then for each p and q there is a path from $a_{p,j1} \neq 0$, $a_{j1,j2} \neq 0$, $\ldots$, $a_{j_r,q} \neq 0$.*

Theorem 6.1. *Let A be an $m \times m$ matrix with associated eigenvalue and eigenvector $Ax = \lambda x$ with $\|x\|_\infty = 1$. Define $D_i := \{z \ : \ |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\}$ for $i = 1, 2, \ldots, m$ to be the Gershgorin Discs. Then $\lambda \in D := \bigcup_{i=1}^m D_i$. Moreover, if A is irreducible then if λ is an eigenvalue of A on the boundary of D, it must be on the boundary of each D_i.*

The first portion of Theorem 6.1 is proven as follows. As $\|x\|_\infty = 1$ there exists an i such that $|x_i| = 1$. Then expanding the i^{th} row of $Ax = \lambda x$ gives $(a_{ii} - \lambda)x_i = \sum_{j \neq i} a_{ij} x_j$. Taking absolute values and bounding the right hand side of the equality using the triangle inequality gives

$$|a_{ii} - \lambda| \leq \sum_{j \neq i} |a_{ij}||x_j|/|x_i| \leq \sum_{j \neq i} |a_{ij}|$$

with the last inequality following from $|x_j| \leq |x_i| = 1$ for all j. Lacking knowledge about which i we have this inequality for we can only ensure that λ is in the union of all such discs. The second portion of Theorem 6.1 follows by noting that if λ is on the border of D then it cannot be on the interior of a disc D_i, so if it is contained in a disc it must be on the boundary of that disc. Once it is known that λ is on the boundary of the i^{th} disc we know that both $|a_{ii} - \lambda| \leq \sum_{j \neq i} |a_{ij}|$ and $|a_{ii} - \lambda| = \sum_{j \neq i} |a_{ij}|$ which is only possible if $|x_j| = 1$ for $j \in \{\ell \ : \ a_{i\ell} \neq 0\}$. Knowing more entries in x where it achieves its max in magnitude allows for the discs of more rows of A to be considered. If A is irreducible this process will continue to all all rows, concluding that $|x_i| = 1$ for all i, and that λ is on the boundary of each disc. This last property is particularly useful for the matrix associated with (31) for Dirichlet problems, which are necessarily IRDD. Theorem 6.1 implies that IRDD matrices are invertible as one of the discs will not contain the origin.

6.2. **Rotated five point stencil.** Poisson's equation (30) was previously approximated (31) using standard symmetric approximations to u_{xx} and u_{yy}. In two dimensions there is greater flexibility in the structure of the stencil, such as by rotating the stencil. For example, note the Taylor series

approximation of $u_{j+1,k+1}$ about the point (x_j, y_k)

$$\begin{aligned} u_{j+1,k+1} &= u + h(u_x + u_y) + \frac{1}{2}h^2(u_{xx} + 2u_{xy} + u_{yy}) \\ &+ \frac{1}{6}h^3(u_{xxx} + 3u_{xxy} + 3u_{xyy} + u_{yyy}) \\ &+ \frac{1}{24}h^4(u_{xxxx} + 4u_{xxxy} + 6u_{xxyy} + 4u_{xyyy} + u_{yyyy}) + \mathcal{O}(h^5) \end{aligned}$$

and

$$\begin{aligned} u_{j+1,k-1} &= u + h(u_x - u_y) + \frac{1}{2}h^2(u_{xx} - 2u_{xy} + u_{yy}) \\ &+ \frac{1}{6}h^3(u_{xxx} - 3u_{xxy} + 3u_{xyy} - u_{yyy}) \\ &+ \frac{1}{24}h^4(u_{xxxx} - 4u_{xxxy} + 6u_{xxyy} - 4u_{xyyy} + u_{yyyy}) + \mathcal{O}(h^5) \end{aligned}$$

where unless otherwise stated u is taken to be at the point (x_j, y_k). From these approximations it is easy to see that

$$\frac{1}{2h^2}(u_{j+1,k+1} + u_{j-1,k-1} + u_{j+1,k-1} + u_{j-1,k+1} - 4u_{j,k}) = \tilde{\tau}_{j,k}$$

where $\tilde{\tau}_{j,k} = \frac{h^2}{12}(u_{xxxx} + 6u_{xxyy} + u_{yyyy}) + \mathcal{O}(h^4)$. Though this finite difference approximation of $u_{xx} + u_{yy}$ differs from that in (31) and they have the same order, it isn't possible to make a combination of them which is of a higher order due to the cross term u_{xxyy} in $\tilde{\tau}_{j,k}$ which isn't involved in the truncation error of the non-rotated five point stencil.

6.3. **Domain boundaries which do not align with equispaced grids.** In this subsection we return to the stencil used in (31). For points (x_j, y_k) which are further than h from the boundary the stencil contains all five points. If the point (x_j, y_k) is a distance h from the boundary $\partial\Omega$ then one or more of the stencil values will be on the boundary, which for Dirichlet boundary conditions will be reflected by the row of the associated matrix having one or more of the non-diagonal entries missing (represented on the right hand side of the linear system); such a row will be strictly diagonally dominant accounting for the matrix being IRDD and invertible as shown in the prior lecture. However, if (x_j, y_k) is closer to a boundary than h in either the x or y direction the approximation in (31) will need to be modified accordingly. Consider for instance a point (x_j, y_k) for which (x_{j+1}, y_k) is not in the interior, but the other stencil values are contained in the interior of Ω. It is then necessary to compute an approximation of u_{xx} from $u_{j-1,k}$, $u_{j,k}$, and $u_{j+\theta,k}$ for some $\theta \in (0, 1)$ corresponding to an approximation at $(x_j + \theta h, y_k)$;

$$\begin{aligned} \alpha u_{j-1,k} + \beta u_{j,k} + \gamma u_{j+\theta,k} &= (\alpha + \beta + \gamma)u_{j,k} \\ &+ (\gamma\theta - \alpha)hu_x + (\gamma\theta^2 + \alpha)\frac{1}{2}h^2u_{xx} \\ &+ (\gamma\theta^3 - \alpha)\frac{1}{6}h^3u_{xxx} + (\mathcal{O})(h^4). \end{aligned}$$

The highest order approximation of u_{xx} is achieved by setting $\alpha + \beta + \gamma = 0$, $\gamma\theta - \alpha = 0$, and $\gamma\theta^2 - \alpha = 2h^{-2}$; giving

$$\begin{aligned} &\frac{u_{j,k+1} + u_{j,k-1} - 2(1 + \theta^{-1})u_{j,k} + 2(1 + \theta)^{-1}u_{j-1,k} + 2\theta^{-1}(1 + \theta)^{-1}u_{j+\theta,k}}{h^2} \\ = \; & u_{xx} + u_{yy} + \frac{1}{12}h^2u_{yyyy} - \frac{1}{3}h(1 - \theta)u_{xxx} + \mathcal{O}(h^2). \end{aligned}$$

Note that the prior stencil and second order accuracy is recovered if θ is equal to one, but reduces to first order in h otherwise; with $\theta \neq 1$ for some point required if the boundary $\partial\Omega$ does not align with the equispaced grid.

In the associated linear system the weighted point $u_{j+\theta,k}$ would be moved to the right hand side of the equation as a known value, resulting in a system that is strictly diagonally dominant with the origin being $2\theta^{-1}(1+\theta)^{-1}h^{-2}$ away from the Gershgorin disc for the associated row of the matrix. This ensures that the system is strictly diagonally dominant for at least one row, and the connected stencil ensures the matrix is irreducible, ensuring that the linear system is invertible.

Part II: NUMERICS FOR CONSERVATION LAWS

7. Lecture 7

Strongly suggested reading (to get acquainted or to recall properties of conservation laws and the nature of their solutions): chapters 1-3 of Leveque's "green" book.

Consider the Cauchy problem in 1D

$$\begin{aligned} \partial_t u + A\partial_x u = 0, \qquad & x \in \Omega = [0, L],\ t \in (0, T] \\ u(x, 0) = u_0(x) & \end{aligned} \tag{33}$$

equipped with suitable BCs.

We proceed to discretise the space-time domain:

$$\begin{aligned} x_j &= jh = \frac{L}{N} j, \quad j = 0, \dots, N \\ t^n &= nk = \frac{T}{K} n, \quad n = 0, \dots, K. \end{aligned}$$

We stick to the case of constant $h = \Delta x$, $k = \Delta t$, but generalisations are not difficult.

Equation (33) has two partial derivatives that can be approximated in many different ways. In the context of finite differences, one can take forward, backward or centred differences for each term, yielding a large number of possible methods:

- One-sided explicit
$$u_j^{n+1} = u_j^n - \frac{k}{h} A(u_{j+1}^n - u_j^n)$$
- One-sided explicit
$$u_j^{n+1} = u_j^n - \frac{k}{h} A(u_j^n - u_{j-1}^n)$$
- Two-step, explicit central scheme
$$\frac{u_j^{n+1} - u_j^{n-1}}{2k} + \frac{A}{2h}(u_{j+1}^n - u_{j-1}^n) = 0$$
- Implicit central (backward Euler) scheme
$$\frac{u_j^{n+1} - u_j^n}{k} + \frac{A}{2h}(u_{j+1}^{n+1} - u_{j-1}^{n+1}) = 0$$
- Explicit central (for $\sigma = 0$ we recover the so-called formwar Euler) scheme
$$\frac{u_j^{n+1} - u_j^n}{k} + \frac{A}{2h}(u_{j+1}^n - u_{j-1}^n) = \frac{\sigma}{h^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n),$$
for $\sigma = \frac{h^2}{2k}$ we get a variant of the Lax-Friedrichs method
$$u_j^{n+1} = \frac{1}{2}(u_{j+1}^n + u_{j-1}^n) - \frac{kA}{2h}(u_{j+1}^n - u_{j-1}^n),$$
and for $\sigma = \frac{kA^2}{2}$ we get the Lax-Wendroff scheme
$$u_j^{n+1} = u_j^n - \frac{kA}{2h}(u_{j+1}^n - u_{j-1}^n) + \frac{k^2A^2}{2h^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n)$$

Remark 7.1. *Recall that a numerical method is **explicit** if $\{u_j^{n+1}\}$ can be computed from $\{u_j^n\}$ directly, for $k \leq n$.*

A scheme is implicit if cannot be recast as explicit.

*A scheme is of **one step** (or also two-level) if it involves only $\{u_j^{n+1}\}$ and $\{u_j^n\}$ (only two time levels).*

Let us now write a "general explicit" scheme for (33) as

$$u_j^{n+1} = G_k(u^n)_j = G(u_{j-1}^n, u_j^n, u_{j+1}^n), \quad \forall j. \tag{34}$$

If (33) has a solution u then we must have

$$u(x_j, t^{n+1}) = G_k(u(t^n))_j - k\tau_j^n,$$

where the last term on the RHS is a measure of the difference between the exact and approximate problems:

For a general two-level method, the **local truncation error** is

$$\tau_j^n = \frac{1}{k}[-u(x_j, t^{n+1}) + G_k(u(\cdot, t^n))_j].$$

The local **total error** is

$$\epsilon_j^{n+1} := u(x_j, t^{n+1}) - u_j^{n+1} = u(x_j, t^{n+1}) - G_k(u^n)_j.$$

Note that at the first time step

$$\begin{aligned}
\epsilon_j^1 &= u(x_j, t^1) - u_j^1, \\
&= -k\tau_j^0 + G_k(u(\cdot, t^0))_j - G_k(u^0)_j \\
&= -k\tau_j^0 + G_k(u(\cdot, t^0) - u^0)_j \\
&= -k\tau_j^0 + G_k(\epsilon^0)_j.
\end{aligned}$$

Applying the same idea for later times, we get

$$\epsilon_j^n = G_k^{(n)}(\epsilon^0)_j - k\sum_{i=0}^{n-1} G_k^{(n-i+1)}(\tau^i)_j, \tag{35}$$

where $G_k^{(n)} = G_k(G_k(\cdots(\cdot)\cdots))$ and $G_k^0 = 1$. This occurs locally. Taking the norm

$$\|v\|_{hp} := \left(h\sum_j v_j^2\right)^{1/2},$$

(where for $p = \infty$ we recover the usual maximum norm), and using triangle inequality in (35) we get

$$\begin{aligned}
\|\epsilon^n\|_{h,p} &\leq \|G_k^{(n)}(\epsilon^0)\|_{hp} + k\sum_{i=0}^{n-1} \|G^{(n-i-1)}(\tau^i)_j\|_{hp} \\
&\leq \|G_k^{(n)}\|hp\|\epsilon^0\|_{hp} + kn\max_i \|G^{(n-i-1)}\|_{hp}\|\tau_j^i\|_{hp},
\end{aligned} \tag{36}$$

where

$$\|G_k^{(n)}\|hp := \sup_{v_j^n \neq 0} \frac{\|G_k^{(n)}(v^n)_j\|_{hp}}{\|v_j^n\|_{hp}}.$$

We made precise the concept of consistency: A scheme with truncation error τ_j^n is **consistent** if

$$\|\tau_j^n\| \to 0 \qquad \text{as } k \to 0,$$

for some norm. Moreover, the consistency is **of order** (s, ℓ) if $\|\tau\| = O(h^s, k^\ell)$. Also, we will say that the operator G induces a **stable method** (aka Lax-Richtmyer stability), if

$$\|G_k^n\|_{hp} \leq C, \qquad \text{for sufficiently small } k.$$

From (36) we see that if (34) is stable, then

$$\|\epsilon^n\|_{hp} \leq CT \max_i \|\tau^i\|_{hp} + C\|\epsilon^0\|_{hp}. \tag{37}$$

If the initial datum has no error associated (or if it has an error that decays with rate $O(k^\ell)$ when $k \to 0$) then

$$\|\epsilon^n\|_{hp} \leq Ck^\ell,$$

which guarantees convergence of the numerical method.

As we have done for the previous lectures, the convergence properties can be established via consistency and stability of the scheme. Let us write the matrix system associated to the finite difference method for (33):

$$\boldsymbol{u}^{n+1} = \mathbf{G}_k \boldsymbol{u}^n,$$

where $\mathbf{G}_k$ is an amplification matrix (since G is a linear operator). A sufficient condition for stability is that

$$\|G_k\|_{hp} \leq 1 + \alpha k. \tag{38}$$

To see this, we can write

$$\|G_k^{(n)}\|_{hp} \leq \|G_k\|_{hp}^n \leq (1 + \alpha k)^n \leq \exp(\alpha n k) = \exp(\alpha T).$$

But in general, establishing that G_k is power-bounded is not a straightforward taks.

8. Lecture 8

Theorem 8.1 (Kreiss)**.** *Assume that A represents a family of matrices. Then the following conditions are equivalent:*

(1) *there exists $C > 0$ such that*
$$\|\mathbf{A}^n\|_{hp} \leq C.$$
(2) *there exists $C > 0$ such that*
$$\frac{1}{\|\lambda\mathbf{I} - \mathbf{A}\|_{hp}} \leq \frac{C}{|\lambda| - 1}, \qquad \forall \lambda \in \mathbb{C}, |\lambda| > 1.$$
(3) *there exists a nonsingular $\mathbf{S}$ and $C > 0$ with $\|\mathbf{S}\|_{hp}\|\mathbf{S}^{-1}\|_{hp} \leq C$, such that $\mathbf{T} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$ is upper triangular and*
   - $|T_{ij}| \leq 1, \quad \forall i,$
   - $|T_{ij}| \leq C \min(1 - |T_{ii}|, 1 - |T_{jj}|), \ i < j.$
(4) *there exists $C > 0$ and $\mathbf{H}$ with $C^{-1}\mathbf{I} \leq \mathbf{H} \leq C\mathbf{I}$, such that $\mathbf{A}^T\mathbf{H}\mathbf{A} \leq \mathbf{H}$.*

Back to our numerical scheme: If $\mathbf{G}_k$ is is uniformly diagonalisable[1], then the Cayley-Hamilton theorem gives
$$\|\mathbf{G}_k^n\|_{hp} \leq 1 \Leftrightarrow \max_i |\Lambda_{ii}| \leq 1.$$

This indicates the Von-Neumann stability criterion: if $\mathbf{G}_k$ is diagonalisable for all k, h, then stability holds whenever $\max_i |\Lambda_{ii}| \leq 1$.

Example. Let us consider $A > 0$ in (33) and apply periodic BCs
$$u(0, t) = u(L, t).$$
The discretisation of the problem can be carried out with the one-sided explicit scheme, giving
$$u_j^{n+1} = u_j^n - \frac{k}{h}A(u_j^n - u_{j-1}^n),$$
for all j associated to interior points. In matrix form, the FD method reads
$$\boldsymbol{u}^{n+1} = \mathbf{G}_k\boldsymbol{u}^n, \qquad \text{with } \mathbf{G}_k = \begin{pmatrix} 1-\lambda & 0 & \cdots & \lambda \\ \lambda & 1-\lambda & \cdots & \\ \vdots & \cdots & & \\ 0 & \cdots & \lambda & 1-\lambda \end{pmatrix},$$
and $\lambda = \frac{Ak}{h}$.

The matrix $\mathbf{G}_k$ is normal (one can readily check that $\mathbf{G}_k^T\mathbf{G}_k = \mathbf{G}_k\mathbf{G}_k^T$) and therefore is diagonalisable. In addition, its eigenvalues are the m roots of
$$\underbrace{\left(\frac{1 - \lambda - \mu}{-\lambda}\right)}_{=:y}{}^m = 1.$$
Then $y^m = 1$ implies that $y_l = \exp(i\frac{2\pi}{m}l)$ for $l = 0, \ldots, m-1$, and thus $\mu_l = 1 - \lambda + \lambda\exp(i\frac{2\pi}{m}l)$ for $l = 0, \ldots, m-1$.

Real eigenvalues: $\mu_0 = 1 - 2\lambda$ and $\mu_{m/2} = 1$. Their absolute value is bounded by 1 if $\lambda \leq 1$. Therefore we have stability (in the Von-Neumann sense) if $\frac{Ak}{h} \leq 1$.

[1]There exists $\mathbf{S}$ with $\|\mathbf{S}\|\|\mathbf{S}^{-1}\| \leq C$ such that $\Lambda = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$ is diagonal.

Remark 8.1. *If $A < 0$, then a similar reasoning as above gives that stability is not achieved.*

Another way is using the so-called Von-Neumann analysis (cf. any textbook).

Let us consider the problem

$$\begin{aligned} \partial_t u + a\partial_x u = 0 \qquad & a > 0,\ x \in (0, L) \\ u(x, 0) = u_0(x), & \\ u(0, t) = u(L, t), & \end{aligned}$$

with $L = 2\pi$, and define the explicit central scheme

$$u_j^{n+1} = u_j^n - \frac{ak}{2h}(u_{j+1}^n - u_{j-1}^n) + \frac{\sigma k}{h^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n).$$

We proceed to seek solutions of the form

$$u_j^n = \hat{u}_l^n \exp(ilx_j).$$

Inserting this in the scheme will provide an amplification coefficient

$$\hat{G}_k = 1 + i\underbrace{\lambda}_{ak/h}\sin(lh) - 4\underbrace{\nu}_{\sigma k/h^2}\underbrace{\beta}_{\sin^2(\frac{lh}{2})},$$

so that the condition $\hat{G}_k \leq 1$ will lead to

$$(16\nu^2 - 4\lambda^2)\beta^2 + (4\lambda^2 - 8\nu)\beta \leq 0. \tag{39}$$

Inspection of each possible case implies that the worst case scenario occurs when $\nu = 1/2$ and so the condition for stability is $0 < \lambda \leq 1$.

Some remarks are in order:

- For the Lax-Friedrichs scheme we had $\sigma = h^2/(2k)$. Therefore $\nu = 1/2 \geq \lambda/2$ and stability holds for $\lambda \leq 1$.
- For Lax-Wendroff: $\sigma = a^2k/2$, giving $\nu \leq \lambda/2$ provided that $\lambda \leq 1$.
- For systems of p PDEs:

$$\begin{aligned} \partial_t \boldsymbol{u} + \mathbf{A}\boldsymbol{u} = \mathbf{0} \\ +\text{BCs} + \text{ICs} \end{aligned}$$

with $\boldsymbol{u}(u_1, \ldots, u_p)$, the stability condition extends for all eigenvalues of $\mathbf{A}$:

$$\frac{k}{h}|\mu_i(\mathbf{A})| \leq 1, \qquad \text{for all real eigenvalues } \mu_i \text{ of } \mathbf{A}.$$

The CFL condition. For the first order PDE above we encountered the condition

$$a\frac{k}{h} \leq 1, \tag{40}$$

to ensure stability. Notice that a is the velocity of propagation of the wave, and we can define a *numerical speed of propagation*

$$v_{num} = \frac{h}{k} \geq a \tag{41}$$

that has to be larger than the physical one, to respect stability.

Definition 8.1. *The domain of dependence is the set of points for which $u_0(x)$ could affect the solution at (x, t).*

Then, stability can be regarded as a condition on the domain of dependences: that of the numerical scheme must contain the one of the physical problem. Condition (41) is the so-called CFL condition (Courant-Friedrichs-Levy, 1928), and we anticipate that it is only a necessary condition for stability.

9. Lecture 9

Upwind method. Recall that for (33) with $A > 0$, the one-sided method

$$u_j^{n+1} = u_j^n - \frac{k}{h} A(u_j^n - u_{j-1}^n),$$

is stable if $0 \leq \frac{Ak}{h} \leq 1$. Notice that the stencil points in the upstream or **upwind** direction.

If $A < 0$ then

$$u_j^{n+1} = u_j^n - \frac{k}{h} A(u_{j+1}^n - u_j^n)$$

is the upwind method.

Conservative methods.

The solution of the linear Riemann problem

$$\begin{cases} u_t + Au_x &= 0 \qquad x \in \mathbb{R},\ t \geq 0 \\ u_0(x) &= \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases}, \end{cases} \tag{42}$$

is the wave $u_0(x - At)$, and the accuracy of its numerical approximation is will definitely be affected, specially near the disctontinuity. The expressions found for the truncation errors (and ruling the rate of convergence of finite difference discretisations), will depend on the same regularity assumptions as Taylor's expansion.

Note that

- First order methods (LF and upwind) give smeared solutions whereas second order methods (LW and BW) capture better the location of the discontinuity but produce oscillations.
- Refinement of the mesh does not remove these issues.
- In the nonlinear case we can get nonlinear instability.
- A method converging to the correct solution in the linear case, may not converge to the *correct* solution in the nonlinear case.

Burgers' equation. Let us consider

$$u_t + (\frac{1}{2}u^2)_x = 0, \qquad x \in [0, 1] \tag{43}$$

endowed with a) a smooth initial datum or b) a jump discontinuity as in (42). As $(\frac{1}{2}u^2)_x = uu_x$, then (43) can be written as

$$u_t + uu_x = 0,$$

and the analogy with the linear case indicates that "$A = u$ is the speed of propagation" of the wave. A first attempt to generalise the ideas above suggests to use upwind methods (under the assumption that $u_j^n \geq 0$ for all j, n):

$$u_j^{n+1} = u_j^n - \frac{k}{h} u_j^n (u_j^n - u_{j-1}^n). \tag{44}$$

Alternatively, we can approximate the flux function $\frac{1}{2}u^2$ with a one-sided difference, giving

$$u_j^{n+1} = u_j^n - \frac{k}{2h}((u_j^n)^2 - (u_{j-1}^n)^2). \tag{45}$$

Depending on the initial datum, the solutions generated with either (44) or (45) produce different results. For instance, for an initial discontinuity, scheme (44) gives a wave moving at incorrect speed. On the other hand, for smooth initial data both schemes produce the same solution.

More generally, let us introduce a

Nonlinear conservation law in conservative form.

$$u_t + (f(u))_x = 0, \tag{46}$$

on some domain Ω with given boundary and initial conditions. Let us also define an explicit method of one-step in conservative form:

$$U_j^{n+1} = U_j^n - \frac{k}{h}(F_{j+1/2}^n - F_{j-1/2}^n), \tag{47}$$

with

$$\begin{cases} F_{j+1/2}^n &= F(U_{j-p}^n, \dots, U_{j+q}^n), \\ F_{j-1/2}^n &= F(U_{j-p-1}^n, \dots, U_{j+q-1}^n) \end{cases},$$

and where F is the numerical flux function (with $p+q+1$ arguments). Here U_j^n stands for either the finite difference pointwise approximation

$$u_j^n \approx u(x_j, t^n),$$

or a local cell average (to be re-discussed later on).

The numerical flux F should represent the exact flux function f. More precisely, we have the following

Definition 9.1. *A numerical flux F is consistent with the conservation law (46) if*

$$F(u, u, \dots, u) = f(u).$$

Moreover, F is Lipschitz continuous if there exists $M > 0$ such that

$$|F(U_0, \dots, U_{p+q}) - f(u)| \le M \max(|U_0 - u|, \dots, |U_{p+q} - u|),$$

provided $|U_i - u|$ is sufficiently small.

The presentation will be restricted to the case $p+q = 1$, but other cases can be straightforwardly described.

Theorem 9.1. *A scheme designed to solve the conservation law (46), and given in conservation form with a consistent and Lipschitz continuous numerical flux, is conservative at the discrete level.*

Proof. Let us confine to scheme (47) with $p = 0, q = 1$, and consider the domain $\Omega = (0, L)$. Summing over all j's we obtain

$$h\sum_j U_j^{n+1} - h\sum_j U_j^n = -k\sum_j (F_{j+1/2}^n - F_{j-1/2}^n),$$

and notice that the terms on the RHS telescope and cancel out (except for the terms at the endpoints). The consistency of F implies that

$$F_{N+1/2}^n = f(u(L, t^n)), \quad F_{-1/2}^n = f(u(0, t^n)),$$

so

$$h\sum_j U_j^{n+1} - h\sum_j U_j^n = -k(f(u(L, t^n)) - f(u(0, t^n))). \tag{48}$$

If we assume periodic boundary conditions we get

$$h\sum_j U_j^{n+1} = h\sum_j U_j^n, \qquad \forall n,$$

and therefore

$$h\sum_j U_j^{n+1} = h\sum_j U_j^0. \tag{49}$$

Now, if we suppose that U_j^0 is such that

$$h\sum_j U_j^0 = \int_\Omega u_0(x)\,\mathrm{d}x, \tag{50}$$

then (49) indicates that the mass of u is preserved throughout the computation.

For constant values of the solution outside Ω, we can rewrite (48) as: for $n = 0, \ldots, K-1$ do

$$h\sum_j U_j^{n+1} - h\sum_j U_j^n = -k[f(u(L)) - f(u(0))],$$

and summing on n and using (50) gives

$$\int_\Omega u(x,T)\,\mathrm{d}x - \int_\Omega u(x,0)\,\mathrm{d}x = -T[f(u(L)) - f(u(0))] = \int_0^T [f(0,t) - f(L,t)]\,\mathrm{d}t, \tag{51}$$

which is precisely the *integral form* of (46). □

10. Lecture 10

The function u is a *weak solution* of (46) if

$$\int_0^T \int_\Omega [u\partial_t\phi + f(u)\partial_x\phi]\,\mathrm{d}x\,\mathrm{d}t = -\int_\Omega u_0(x)\phi(x,0)\,\mathrm{d}x, \tag{52}$$

holds for all smooth *test functions* $\phi(x,t)$ (compactly supported in Ω).

Definition 10.1. *The discrete total variation (TV) of a function v is*

$$TV(v) := \sup \sum_{j=1}^{N} |v(\xi_j) - v(\xi_{j-1})|,$$

where the supremum is taken over all possible subdivisions of $\mathbb{R}$.

Definition 10.2. *A function v is* total-variation-bounded *if $TV(v) < \infty$.*

Definition 10.3. *The discrete solution generated with a numerical scheme converges to a TV-bounded function in L_1 if*

$$\|U_\ell - u\|_{1,\Omega} := \int_0^T \int_\Omega |U_\ell(x,t) - u(x,t)|\,\mathrm{d}x\,\mathrm{d}t \to 0, \qquad \text{when } \ell \to \infty,$$

where ℓ denotes refinement of both space and time discretisations h^ℓ, k^ℓ.

Theorem 10.1 (Lax-Wendroff)**.** *Suppose that a numerical scheme is in conservative form and it has a consistent and Lipschitz continuous numerical flux. If the method converges to a TV-bounded function, then that function is indeed a weak solution of the continuous conservation law.*

Proof. Again, let us focus on $p = 0, q = 1$ and write a *discrete weak form* for (47). That is, we "multiply by a test function and integrate over the domain", but in a discrete setting:

$$\sum_{j,n} \frac{U_j^{n+1} - U_j^n}{k}\phi_j^n = -\sum_{j,n} \frac{F_{j+1/2}^n - F_{j-1/2}^n}{h}\phi_j^n,$$

where we also require ϕ to be compactly supported in Ω (implying that $\phi(x_j, t^n) = 0$ for $j \to \pm\infty, n \to \infty$).

Therefore, using the identity (representing the telescope property together with a discrete integration by parts)

$$\sum_{n=0}^{p} a^n(b^{n+1} - b^n) = -a^0b^0 - \sum_{n=1}^{p} b^n(a^n - a^{n-1}) + a^pb^{p+1}$$

for $a = \phi$ and $b = U$ (in combination with the fact that ϕ has compact support in Ω) we obtain

$$\sum_{j,n} U_j^n \frac{\phi_j^n - \phi_j^{n-1}}{k} - \frac{1}{k}\sum_j \phi_j^0 U_j^0 = \sum_{j,n} F_{j+1/2}^n \frac{\phi_{j+1}^n - \phi_j^n}{h}. \tag{53}$$

Let us now consider a family of discretisations such that $(h^\ell, k^\ell) \to 0$ as $\ell \to \infty$, and multiply (53) by $h^\ell k^\ell$. The assumption of L_1 convergence for U_j^n with the smoothness of ϕ lead to

$$\lim_{\ell\to\infty} \sum_{j,n} U_j^n \frac{\phi_j^n - \phi_j^{n-1}}{k^\ell} h^\ell k^\ell = \int_0^T \int_\Omega u \partial_t \phi \, \mathrm{d}x \, \mathrm{d}t.$$

Similarly, the second term on the LHS of (53) converges

$$\lim_{\ell\to\infty} h^\ell \sum_j \phi_j^0 U_j^0 = \int_\Omega u_0(x)\phi(x,0) \, \mathrm{d}x.$$

For the remainder of (53) we use the TV-boundedness of U^n together with the consistency and Lipschitz continuity of F to arrive at

$$\lim_{\ell\to\infty} h^\ell k^\ell \sum_{j,n} F(U_j^n, U_{j+1}^n) \frac{\phi_{j+1}^n - \phi_j^n}{h^\ell} = \int_0^T \int_\Omega f(u) \partial_x \phi \, \mathrm{d}x \, \mathrm{d}t.$$

Consequently, if U_j^n converges then it converges to the solution of the weak form of (52). □

Entropy solutions. (Suggested reading: Section 3.8 of LeVeque's book). Recall that weak solutions might not be unique and an additional condition is required to pick the *physically relevant* solution. Depending on the properties of the flux function, different characterisations of the entropy condition are available.

Definition 10.4. *A numerical method of the form*

$$U_j^{n+1} = U_j^n - \lambda(F_{j+1/2}^n - F_{j-1/2}^n) = G(U_{j-p-1}^n, \ldots, U_{j+q}^n) = G(U)_j^n,$$

is said monotone *if it is monotonically non-decreasing in all its arguments.*

For example, restricting again to $p = 0, q = 1$ we put $G(U^n)_j = U_j^n - \lambda(F_{j+1/2}^n - F_{j-1/2}^n) = G(U_{j-1}^n, U_j^n, U_{j+1}^n)$ and verify that

$$\frac{\partial G}{\partial U_{j-1}^n} = \lambda \frac{\partial}{\partial U_{j-1}^n} F(U_{j-1}^n, U_j^n), \quad \frac{\partial G}{\partial U_j^n} = 1 - \lambda\left(\frac{\partial}{\partial U_j^n} F(U_j^n, U_{j+1}^n) - \frac{\partial}{\partial U_j^n} F(U_{j-1}^n, U_j^n)\right),$$

$$\frac{\partial G}{\partial U_{j+1}^n} = -\lambda \frac{\partial}{\partial U_{j+1}^n} F(U_j^n, U_{j+1}^n).$$

Then, if F is increasing in its first argument and decreasing in its second argument, and if also λ is sufficiently small; then G is monotone.

Theorem 10.2. *For a monotone scheme we have that*

(1) *If $U_j \le V_j$ then $G(U)_j \le G(V)_j$*
(2) *The produced solution satisfies*

$$\min_{i\in S_j} U_i^n \le G(U^n)_j \le \max_{i\in S_j} U_i^n.$$

(3) *The scheme is L_1-contractive*

$$\|G(U^n)_j - G(V^n)_j\|_1 \le \|U_j^n - V_j^n\|_1.$$

(4) *The method is Total-Variation Diminishing (TVD)*

$$TV(U^{n+1}) \le TV(U^n).$$

11. Lecture 11

Proof. (of theorem 10.2.) Properties (1)-(4) are part of exercises in QS5.

Using the notation $c^+ = \max(c, 0)$, $c^- = \min(c, 0)$, we let $W_j^n = \max(U_j^n, V_j^n) = V_j^n + (U_j^n - V_j^n)^+$ and suppose that statement (1) is valid. Then

$$G(U^n)_j \leq G(W^n)_j, \qquad G(V^n)_j \leq G(W^n)_j,$$

and $W_j^n - V_j^n = (U_j^n - V_j^n)^+$. These relations imply that

$$G(W^n)_j - G(V^n)_j \geq \max(0, G(U^n)_j - G(V^n)_j) = (G(U^n)_j - G(V^n)_j)^+.$$

Summing over all j's and using the conservation property together with the previous considerations, we obtain

$$\sum_j (G(U^n)_j - G(V^n)_j)^+ \leq \sum_j W_j^{n+1} - V_j^{n+1} = \sum_j (U_j^n - V_j^n)^+.$$

Then

$$\sum_j |G(U^n)_j - G(V^n)_j| \leq \sum_j (U_j^n - V_j^n)^+ + \sum_j (V_j^n - U_j^n)^+ = \sum_j |U_j^n - V_j^n|.$$

□

These properties indicate that for monotone schemes we can ensure stability, the absence of oscillations, and essential features of the conservation law will be preserved. Some examples of monotone methods:

- Extension of the Lax-Friedrichs scheme using the Russanov flux:
$$F_{LF}(u, v) = \frac{f(u) + f(v)}{2} - \frac{\alpha}{2}(v - u),$$
with $\alpha = \max_u |f'(u)|$. It is clearly a consistent numerical flux, and we can check that it is "$F_{LF}(\uparrow, \downarrow)$" (increasing and decreasing in its first and second arguments, respectively).
- Extension of the Lax-Wendroff scheme to the nonlinear case:
$$F_{LW}(u, v) = \frac{f(u) + f(v)}{2} + \frac{\lambda}{2} f'(\frac{u + v}{2})(f(v) - f(u)).$$
It is consistent, but taking as an example the rescaled Burgers flux $f(u) = u^2$ we can see that the numerical flux is not is not increasing in its first argument and therefore not monotone.
- The upstream flux:
$$F_U(u, v) = \begin{cases} f(u), & s \geq 0 \\ f(v), & s < 0 \end{cases},$$
where s is the shock speed based on the Rankine-Hugoniot condition at (u, v). It is clearly consistent, but again, the monotonicity cannot be verified for any values of u, v.

Let us now resume our previous discussion about entropy solutions, and we recall that additional conditions are required to identify unique weak solutions.

Entropy conditions.

Definition 11.1. *A discontinuity propagating with speed* $s = \frac{f(u_R) - f(u_L)}{u_r - u_L}$ *satisfies an* entropy condition *if*

$$f'(u_L) > s > f'(u_R).$$

If f *is convex, this condition translates in* $u_L > u_R$.

Definition 11.2 (Oleinik)**.** *For non-convex fluxes, the previous definition extends to*

$$\frac{f(u)-f(u_L)}{u-u_L} \geq s \geq \frac{f(u)-f(u_R)}{u-u_R},$$

for all u between u_L and u_R.

Definition 11.3. *For non-shock solutions (e.g. rarefaction waves): u is the entropy solution if there exists $E > 0$ such that for all $a > 0$, $t > 0$, $x \in \mathbb{R}$:*

$$\frac{u(x+a,t)-u(x,t)}{a} < \frac{E}{t}.$$

For discontinuous solutions we recover the first definition.

Definition 11.4 (General case)**.** *The function u is the entropy solution if for all convex entropy functions and corresponding entropic fluxes (η, ψ) there holds*

$$\int_0^T \int_\Omega \partial_t \phi \eta(u) + \partial_x \phi \psi(u) \,\mathrm{d}x \,\mathrm{d}t \geq -\int_\Omega \phi(x,0)\eta(u(x,0)) \,\mathrm{d}x,$$

for all non-negative test functions ϕ.

This definition requires (η, ψ) to be an *entropy pair*, that is, the convex function and entropic flux must satisfy

$$\eta(u)_t + \psi(u)_x = 0, \quad \text{and} \quad \psi'(u) = \eta'(u)f'(u),$$

and so $\psi(u) = \int_0^u \eta'(v)f'(v)dv$.

A well-known example is the Kruzkov entropy pair

$$\eta(u) = |u-c|, \qquad \psi(u) = \operatorname{sgn}(u-c)[f(u)-f(c)],$$

for a real c.

Theorem 11.1. *If the entropy inequality for a given conservation law is satisfied for the Kruzkov pair for all $c \in \mathbb{R}$, then it is also satisfied for any entropy pair.*

Theorem 11.2. *The solution generated by a monotone scheme satisfies all entropy conditions.*

Sketch of a proof. (1) Use the Kruzkov entropic pair for a given $c \in \mathbb{R}$.
(2) Assume that the entropy inequality holds weakly.
(3) Write a local entropy condition for the discrete solution

$$\frac{\eta(U_j^{n+1}) - \eta(U_j^n)}{k} + \frac{\psi_{j+1/2}^n - \psi_{j-1/2}^n}{h} \leq 0,$$

where $\psi^n_{j+1/2} = F(\max(U^n, c))^n_{j+1/2} - F(\min(U^n, c))^n_{j+1/2}$.

(4) Rearranging terms one ends up with

$$G(\max(U^n, c))_j - G(\min(U^n, c))_j = |U^n_j - c| - \lambda(\psi^n_{j+1/2} - \psi^n_{j-1/2}).$$

(5) Applying consistency and monotonicity of the scheme we can show that

$$\max(c, U^{n+1}_j) \leq G(\max(c, U^n))_j.$$

(6) Then we can bound $\eta(U^{n+1}_j)$ using the previous step and the entropy condition from step (3) can be proved locally, implying also a global discrete entropy condition.

□

12. Lecture 12

Definition 12.1. *A numerical scheme is monotonicity-preserving if*

$$U_{j+1}^n \geq U_j^n \ \forall j \Longrightarrow U_{j+1}^{n+1} \geq U_j^{n+1} \ \forall j.$$

Theorem 12.1. *A total variation diminishing (TVD) scheme is monotonicity-preserving.*

Proof. Let us assume that the TVD property holds and that $U_{j+1}^n \geq U_j^n \ \forall j$. On the other hand, suppose that there exists l such that

$$U_{l+1}^{n+1} < U_l^{n+1}.$$

Consider the situation where the solution is constant on the left and on the right of the interval $[x_l, x_{l+1}]$. Then the total variation on each side is zero and when passing from time t^n to t^{n+1} the stencils of U_l and U_{l+1} will interact in such a way that the monotonicity will no longer be preserved. This implies that the TVD property is violated, leading to a contradiction. □

We have the following chain of properties for a numerical scheme:

$$\text{Monotone } \Rightarrow L_1 - \text{contraction } \Rightarrow \text{TVD } \Rightarrow \text{Monotonicity-preserving.}$$

Definition 12.2. *A numerical scheme is linear if it is linear in all its arguments when applied to a linear PDE:*

$$U_j^{n+1} = \sum_{l=-p}^{q} c_l(\lambda) U_{j+l}^n.$$

For linear methods to be monotone we require that $c_l(\lambda) \geq 0$ for all l. They are commonly called *positive schemes*.

Theorem 12.2. *If a linear scheme is monotonicity-preserving, then it is monotone.*

Proof. For a given m, let

$$U_i = \begin{cases} 0 & i \leq m, \\ 1 & i > m \end{cases},$$

which clearly satisfies monotonicity. Then, for a linear scheme

$$U_{j+1}^{n+1} = \sum_{l=-p}^{q} c_l(\lambda) U_{j+l+1}^n, \quad U_j^{n+1} = \sum_{l=-p}^{q} c_l(\lambda) U_{j+l}^n,$$

and so

$$U_{j+1}^{n+1} - U_j^{n+1} = \sum_{l=-p}^{q} c_l(\lambda)(U_{j+l+1}^n - U_{j+l}^n) = \begin{cases} 0 & i \neq m, \\ c_m(\lambda) & i = m \end{cases}.$$

But the LHS is nonnegative since U_i is monotone, which then gives that $c_m(\lambda) \geq 0$. The same reasoning can be done for other m's, implying the monotonicity of the scheme. □

We have another chain of properties for a numerical scheme:

$$\text{Monotonicity-preserving} \underbrace{\Rightarrow}_{\text{for linear schemes}} \text{Monotone} \Rightarrow \text{TVD} \Rightarrow \text{Monotonicity-preserving.}$$

Theorem 12.3 (Godunov)**.** (1) *A linear monotone (TVD) scheme is at most first order accurate.*
(2) *A general monotone scheme is at most first order accurate.*

Proof of (1). Let us suppose the existence of a smooth solution $u(x,t)$ to the conservation law. Then a Taylor expansion on u at some point of the mesh and at some time instant, allows us to write

$$u(x_{j+l}, t^n) = \sum_{r=0}^{\infty} \frac{(lh)^r}{r!} \frac{\partial^r}{\partial x^r} u(x_j, t^n) \approx U_{j+l}^n.$$

Constructing now an approximation of the solution by a linear numerical method we obtain

$$\text{(i)} \qquad U_j^{n+1} := \sum_{l=-p}^{q} c_l(\lambda) \sum_{r=0}^{\infty} \frac{(lh)^r}{r!} \frac{\partial^r}{\partial x^r} u(x_j, t^n).$$

A Taylor expansion can be also applied to the exact solution in terms of the time derivatives:

$$u(x_j, t^{n+1}) = \sum_{r=0}^{\infty} \frac{(k)^r}{r!} \frac{\partial^r}{\partial t^r} u(x_j, t^n).$$

Using now the relation

$$\frac{\partial^r v}{\partial t^r} = (-1)^r \frac{\partial^r v}{\partial x^r},$$

which holds in particular for the linear equation $u_t + u_x = 0$, we can combine the two Taylor expansions above to arrive at

$$\text{(ii)} \qquad u(x_j, t^{n+1}) = \sum_{r=0}^{\infty} (-1)^r \frac{(k)^r}{r!} \frac{\partial^r}{\partial x^r} u(x_j, t^n).$$

For $r = 0$, (i) and (ii) imply, respectively, that

$$U_j^{n+1} \approx \sum_{l=-p}^{q} c_l(\lambda) u(x_j, t^n), \quad u(x_j, t^n) \approx u(x_j, t^{n+1}),$$

and therefore

$$\text{(iii)} \qquad \sum_{l=-p}^{q} c_l(\lambda) = 1.$$

For $r = 1$, (i) and (ii) imply, respectively, that

$$U_j^{n+1} \approx \sum_{l=-p}^{q} c_l(\lambda)[u(x_j, t^n) + lh\partial_x u(x_j, t^n)], \quad u(x_j, t^n) \approx u(x_j, t^{n+1}) - k\partial_x u(x_j, t^n),$$

and therefore

$$\text{(iv)} \qquad \sum_{l=-p}^{q} l c_l(\lambda) = -\lambda.$$

For $r = 2$, (i) and (ii) imply, respectively, that

$$U_j^{n+1} \approx \sum_{l=-p}^{q} c_l(\lambda)[u(x_j, t^n) + lh\partial_x u(x_j, t^n) + \frac{l^2h^2}{2}\partial_{xx} u(x_j, t^n)],$$

$$u(x_j, t^n) \approx u(x_j, t^{n+1}) - k\partial_x u(x_j, t^n) + \frac{k^2}{2}\partial_{xx} u(x_j, t^n),$$

and therefore

$$\text{(v)} \qquad \sum_{l=-p}^{q} l^2 c_l(\lambda) = \lambda^2.$$

The last three relations (iii)-(v) can be put together as

$$1 \cdot \lambda^2 = (-\lambda)^2,$$

and so

$$\big(\sum_{l=-p}^{q} c_l(\lambda)\big)\big(\sum_{l=-p}^{q} l^2 c_l(\lambda)\big) = \big(\sum_{l=-p}^{q} l c_l(\lambda)\big)^2.$$

The monotonicity of the scheme implies that we must have $c_l(\lambda) \geq 0$, so we can write $d_l^2 = c_l(\lambda) \geq 0$. Then

$$\text{(vi)} \qquad \big(\sum_{l=-p}^{q} d_l^2\big)\big(\sum_{l=-p}^{q} l^2 d_l^2\big) = \big(\sum_{l=-p}^{q} l d_l^2\big)^2,$$

and after defining the vectors $\boldsymbol{a} = (d_{-q}, \ldots, d_p)^T$, $\boldsymbol{b} = (-qd_{-q}, \ldots, pd_p)^T$, relation (vi) is rewritten as

$$(\boldsymbol{a} \cdot \boldsymbol{a})(\boldsymbol{b} \cdot \boldsymbol{b}) = (\boldsymbol{a} \cdot \boldsymbol{b})^2,$$

which can only occur if $\boldsymbol{a} = C\boldsymbol{b}$ with C constant, which clearly is not possible. This implies that we cannot match the first three terms in the two Taylor expansions above, indicating that linear schemes will produce truncation errors of order at most one.

The proof of statement (2) follows very much in the same way. □

13. Lecture 13

The proof of the last theorem relies on the regularity of the exact solution. However, as we have seen during this part of the course, the solutions of conservation laws will often be discontinuous. The following results address the extension to more general regularity assumptions.

Theorem 13.1 (Tang & Feng)**.** *Assume that the finite difference scheme*

$$U_j^{n+1} = \sum_{l=-p}^{q} c_l(\lambda) U_j^n$$

is has a consistent numerical flux and it is monotone when applied to a linear conservation law

$$u_t + Au_x = 0.$$

Then, for any $M > 0$

$$C(p,q)M \sum_{l=-p}^{q} \sqrt{c_l(\lambda)} \sqrt{\frac{ht^n}{\lambda}} \|U_0\|_{TV} \leq \|u(\cdot,t^n) - U^n\|_{h,1} \leq M \left[2 \sqrt{\sum_{l=-p}^{q} l^2 c_l(\lambda) - \lambda^2 A^2} \sqrt{\frac{ht^n}{\lambda}} + h \right],$$

where $\|U_0\|_{TV} = \sup_h \|U_j^n\|_{h,1}$.

Theorem 13.2 (Kuznetsov)**.** *The numerical solution to the general conservation law*

$$u_t + f(u)_x = 0$$

generated by a scheme being monotone, conservative, and having consistent and Lipschitz continuous numerical flux, converges to the entropy solution, for any initial condition. Moreover, if $\|U_0\|_{TV}$ *is bounded, then*

$$\|u(\cdot,t^n) - U^n\|_{h,1} \leq C(t^n)\sqrt{h}.$$

Finite volume schemes. Considering again the generation of meshes used in the finite difference approximation of one-dimensional problems, let us define cells or control volumes centred at each point of the mesh, and having a size of h, and occupying a time interval of size k:

$$[x_{j-1/2}, x_{j+1/2}] \times [t^n, t^{n+1}],$$

for all j.

Let us recall the general conservation law written in the form

$$u_t + f(u)_x = 0, \tag{60}$$

endowed with appropriate boundary and initial conditions, for which we can state an *integral form* (on the cell centred at x_j):

$$\int_{x_{j-1/2}}^{x_{j+1/2}} (u(x,t^{n+1}) - u(x,t^n))\,\mathrm{d}x = -\int_{t^n}^{t^{n+1}} (f(u(x_{j+1/2},t)) - f(u(x_{j-1/2},t)))\,\mathrm{d}t. \tag{61}$$

Defining cell averages

$$\bar{u}_j^n = \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x,t^{n+1})\,\mathrm{d}x$$

and boundary fluxes

$$F^n_{j+1/2} = F(\bar{u}^n_j, \bar{u}^n_{j+1}) = \frac{1}{k}\int_{t^n}^{t^{n+1}} f(u(x_{j+1/2}, t))\,\mathrm{d}t, \tag{62}$$

one recovers the explicit numerical method in *finite volume (FV) formulation*:

$$\bar{u}^{n+1}_j = \bar{u}^n_j - \frac{k}{h}[F(\bar{u}^n_j, \bar{u}^n_{j+1}) - F(\bar{u}_{-1}j^n, \bar{u}^n_j)].$$

Godunov's method.

Notice that a numerical method (finite difference or finite volume) written in conservation form will capture correctly the wave speed of the discontinuity. However the FV method above is still not completely defined, as the integral characterising the boundary fluxes (62) has to be approximated.

Evaluating this integral requires the point-wise values of the approximated solution at $x_{j+1/2}$ and $x_{j-1/2}$ (which are not points in the initial mesh). Moreover $\bar{u}$ by definition is not necessarily continuous at these points. We then proceed to interpolate:

- Let us assume that $u^* = u(x_{j+1/2}, t)$ can be *reconstructed* from the average values
$$(\bar{u}^n_{j-p}, \dots, \bar{u}^n_{j+q})$$
by some polynomial $p(x)$ sampling u at $x_{j+1/2}$.
- Ideally p should not violate the conservation property. Therefore
$$\int_{x_{l-1/2}}^{x_{l+1/2}} p(x)\,\mathrm{d}x = \bar{u}^n_l, \qquad l \in \{j-p, \dots, j+p\}.$$
- A straightforward example is $u^* = \frac{1}{2}(\bar{u}^n_j + \bar{u}^n_{j+1})$, but higher order reconstructions will be preferable in many situations.
- In general we want to find which value to assign to the approximate solution at a discontinuity. And this will occur at every interface between two cells, that is at $x_{j+1/2}$. This accounts to find the solution of a Riemann problem, locally.
$$u^L = \lim_{\epsilon\to 0^+} p_L(x_{j+1/2} - \epsilon), \qquad u^R = \lim_{\epsilon\to 0^+} p_R(x_{j+1/2} + \epsilon).$$
For sufficiently small k, we observe that u^* will remain constant over $[t^n, t^{n+1}]$, so even the fastest wave will not reach a neighbouring interface within one time step.

In summary, in (62) we will employ

$$F^n_{j+1/2} = f((u^*_{j+1/2}),$$

with u^* reconstructed locally from $\bar{u}^n_j$ and $\bar{u}^n_{j+1}$ (ie. *the exact flux* evaluated on the local Riemann solution). The Godunov numerical flux can be therefore defined as

$$F(u,v) = \begin{cases} f(u) & u \geq v, f(u) \geq f(v) \\ f(v) & u \geq v, f(u) < f(v) \\ f(u) & u < v, f'(u) \geq 0 \\ f(v) & u < v, f'(v) < 0 \\ f(f^{-1}(0)) & \text{otherwise.} \end{cases}$$

If f happens to be convex, then we recover

$$F(u,v) = \begin{cases} \min\limits_{u\leq w\leq v} f(w) & \text{if } u \leq v \\ \max\limits_{v\leq w\leq u} f(w) & \text{if } u > v. \end{cases}$$

A direct inspection of each case indicates that the Godunov numerical flux is indeed monotone. But we still have not specified how we actually compute u^*.

14. Lecture 14

Approximate Riemann solvers. Some examples are provided in what follows.

Local Lax-Friedrichs (Rusanov) numerical flux.

$$F(\bar{u}_j^n, \bar{u}_{j+1}^n) = \frac{1}{2}[f(\bar{u}_j^n) + f(\bar{u}_{j+1}^n) - \alpha_{j+1/2}(\bar{u}_{j+1}^n - \bar{u}_j^n)], \tag{63}$$

Taking the viscosity parameter $\alpha_{j+1/2} = \max_w |f'(w)|$ as the largest local wave speed, the stability of the scheme is guaranteed. Notice that this additional viscosity is small in regions where the solution is smooth.

Exercise: check that this flux is monotone.

Roe numerical flux. It is based on a linearisation of the flux around the cell interface, combined with an upwind approximation. The linear part should approximate the speed suggested by the exact solution of the Riemann problem. For example, taking the Rankine-Hugoniot condition (provided that we do have a jump) we can set

$$a(u_L, u_R) = \frac{f(u_L) - f(u_R)}{u_L - u_R}, \tag{64}$$

otherwise we can use $a(u_L, u_R) = f'(w)$ with $w = u_L$ or $w = u_R$.

Related to (64), notice that

- If $u_R \approx u_L$ then $a(u_L, u_R)$ is a reasonable approximation for both $f'(u_L)$ and $f'(u_R)$.
- If $f'(u_L)$ and $f'(u_R)$ have the same sign, then a will have this sign as well.
- If $f'(u_L)$ and $f'(u_R)$ have different signs (as in e.g. rarefaction waves), then the solution generated using a will differ from the one obtained with an exact Riemann solver (and therefore it might not be the entropy solution).

If $u_L = u_R$, the numerical flux is trivial. Let us take $u_L \neq u_R$ and define

$$F(u_L, u_R) = \begin{cases} f(u_L) & a(u_L, u_R) > 0, \\ f(u_R) & a(u_L, u_R) \leq 0. \end{cases} \tag{65}$$

Using (64) and (65) we get that if $a(u_L, u_R) > 0$ then

$$F(u_L, u_R) = f(u_L) = \frac{1}{2}[f(u_L) + f(u_R) - \frac{f(u_R) - f(u_L)}{u_R - u_L}(u_R - u_L)],$$

and if $a(u_L, u_R) < 0$,

$$F(u_L, u_R) = f(u_R) = \frac{1}{2}[f(u_L) + f(u_R) - \frac{f(u_R) - f(u_L)}{u_R - u_L}(u_R - u_L)].$$

Therefore (65) gives

$$F^{\text{Roe}}(u_L, u_R) = \frac{1}{2}[f(u_L) + f(u_R) - |a(u_L, u_R)|(u_R - u_L)].$$

Exercise: check that this flux is not monotone.

Harten's entropy fix. As mentioned above, a drawback of Roe's numerical flux is that it produces only shock solutions and will not capture e.g. rarefaction waves. Smoothing the graph of the wave speed in Roe's flux remediates this issue:

$$F^{\text{fix}}(u_L, u_R) = \frac{1}{2}[f(u_L) + f(u_R) - \Phi(a(u_L, u_R))(u_R - u_L)],$$

where

$$\Phi(s) = \begin{cases} |s| & |s| \geq \epsilon, \\ \frac{s^2+\epsilon^2}{2\epsilon} & |s| < \epsilon, \end{cases}$$

for some $\epsilon > 0$.

Engquist-Osher numerical flux. Under the assumption that there exists f^+, f^- such that $f(u) = f^+(u) + f^-(u)$ and $\partial_u(f^\pm(u)) = a^\pm(u)$, one has that

$$\int_{u_L}^{u_R} a^\pm(u)du = f^\pm(u_R) - f^\pm(u_L),$$

and rearranging terms we end up with

$$F^{\mathrm{EO}}(u_L, u_R) = \frac{1}{2}[f(u_L) + f(u_R)] - \frac{1}{2}\int_{u_L}^{u_R} |f'(\theta)|d\theta. \tag{66}$$

If f has a single minimum at w and no maximum, we can recast (66) as

$$F^{\mathrm{EO}}(u_L, u_R) = f(\max(u_L, w)) + f(\min(u_R, w)) - f(w),$$

and if f is convex then we can write

$$F^{\mathrm{EO}}(u_L, u_R) = f^+(u_L) + f^-(u_R).$$

Exercise: check wether this flux is monotone.

15. Lecture 15

High order methods. High-order methods would achieve higher rates of convergence, provided that the exact solution is smooth enough. However in the context of hyperbolic conservation laws, the functions involved are typically discontinuous and the formal order of convergences that one could derive will be quite low (recall theorems from Lecture 13). Nevertheless, high order methods (having consistency errors of order at least two) are desirable as for a fixed spatial resolution, they deliver smaller errors than first-order approximations. In addition, they provide better capturing of local features.

Let us consider the following *semi-discretisation* of (60), written in conservation form

$$\partial_t U_j + \frac{F_{j+1/2} - F_{j-1/2}}{h} = 0, \tag{70}$$

and recall that in finite difference methods, $U_j(t)$ will represent the point-wise approximation of $u(x_j, t)$ and the numerical flux $F_{j+1/2} = F(U_{j-p}, \dots, U_{j+q})$ is a *direct* approximation of $f(u)$ evaluated at $x_{j+1/2}$. On the other hand, in finite volume schemes, $U_j(t)$ represents the cell average on the interval I_j and the numerical flux $F_{j+1/2}$ is obtained in two stages: a) recovering $u_{j+1/2}$ from the cell averages $(\bar{u}_{j-p}, \dots, \bar{u}_{j+q})$, and b) evaluating $f(u_{j+1/2})$.

Independently of the specific form of the method at hand, we require a conservation property:

$$f(u(x)) = \int_{x-h/2}^{x+h/2} F(x)\,\mathrm{d}x. \tag{71}$$

Taking derivatives we obtain

$$\partial_x f = \frac{F(x+h/2) - F(x-h/2)}{h}.$$

In turn, the definition of cell averages applied in (71) gives

$$\bar{F}_j = \frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} F(x)\,\mathrm{d}x = f(u_j),$$

which implies that, in finite difference methods, the numerical flux $F_{j+1/2}$ is recovered (from cell averages of F) as the point value of f; whereas in finite volumes, the solution at the interface $u_{j+1/2}$ is recovered (from cell averages of u) and the numerical flux is $f(u_{j+1/2})$.

Let us now consider a numerical method of $m-$th order accuracy. Then, in particular,

$$\frac{F_{j+1/2} - F_{j-1/2}}{h} = \partial_x f|_{x_j} + O(h^m). \tag{72}$$

Remark 15.1. *Equation (72) implies that the construction of high-order methods will translate in reconstructing (with high accuracy) the numerical flux at each interface.*

In the light of Godunov's theorem, *to achieve a high-order discretisation we will require nonlinear methods.*

Let us construct a general method whose numerical flux is recast in the following form

$$F^n_{j\pm1/2} = F^{n,\mathrm{low}}_{j\pm1/2} + \Phi_{j\pm1/2} A^n_{j\pm1/2},$$

where the first term in the RHS indicates a flux of a low-order method (which will be monotone and therefore non-oscillatory). We proceed to write the method as

$$F^n_{j\pm1/2} = F^{n,\mathrm{low}}_{j\pm1/2} + \Phi_{j\pm1/2}[F^{n,\mathrm{high}}_{j\pm1/2} - F^{n,\mathrm{low}}_{j\pm1/2}],$$

where the term in the RHS having a $^{\mathrm{high}}$ superscript denotes the contribution coming from a high-order method satisfying (72), but being possibly oscillatory; and the function Φ stands for a

so-called *flux limiter*, applied locally at the interfaces. A family of different methods can be defined depending on how we choose this flux limiter. Some of these are outlined in what follows.

Flux-correction transport schemes.

Here Φ is chosen such that no extrema are created and existing ones are not accentuated. An example is given by the classical min-mod limiter:

$$\Phi_{j\pm1/2}A^n_{j\pm1/2} = \text{minmod}\,(A^n_{j+1/2}, U^n_{j+1} - U^n_j), U^n_j - U^n_{j-1}),$$

where the min-mod function is defined as

$$\text{minmod}\,(a_1, \ldots, a_n) := \begin{cases} \text{sign}(a_1)\min_i |a_i|, & \sum_i \text{sign}(a_i) = n \\ 0, & \text{otherwise.} \end{cases}$$

TVD-stable and high-order schemes.

This second class of methods is a generalisation to the case where Φ itself is nonlinear, and the idea is based on introducing a smoothness monitor r

$$\Phi_{j+1/2} = \Phi(r_{j+1/2}), \qquad r_{j+1/2} := \frac{U_j - U_{j-1}}{U_{j+1} - U_j},$$

and one notices that $\Phi(1) = 1$ will ensure second order accuracy. The specific form of Φ gives rise to methods with different properties. For example,

- $\Phi(r) = \max(0, \min(r, \beta))$ for some $\beta \leq 1$ (Chakravarty-Osher limiter)
- $\Phi(r) = \max(0, \min(2r, 3(2 + r), 2))$ (Monotonised centred limiter)
- $\Phi(r) = \max(0, \min(r\beta, 1), \min(r, \beta))$ for some $1 \leq \beta \leq 2$. For $\beta = 2$ the method is known as SUPERBEE.
- $\Phi(r) = \frac{3r(r+1)}{2(r^2+r+1)}$ (Ospre)
- $\Phi(r) = \max(0, \frac{2r}{1+r})$ (Van Leer)
- $\Phi(r) = \text{minmod}, (1, r)$

All of these limiters produce methods that are of second order in regions where the solution is smooth, and boil down to the upwind method (monotone and first order) near the discontinuities.

We can also consider the following class of *linear* methods

- $\Phi(r) = 0$ (upwind method, TVD)
- $\Phi(r) = 1$ (Lax Wendroff method, not TVD)
- $\Phi(r) = r$ (Beam warming)
- $\Phi(r) = \frac{1+r}{2}$ (Fromm)

Different kinds of waves will require diverse specific properties of the method, which will have to balance shock capturing capabilities with the smearing of oscillations. The goal will be to choose a limiter as close to 1 as possible, but still enforcing a TVD property.

16. Lecture 16

Example. Let us considet the linear conservation law (setting $f(u) = au$, with a constant; and let us recall the Lax-Wendroff scheme, here specifically written in finite volume formulation

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{\lambda a}{2}(\bar{u}_{j+1}^n - \bar{u}_{j-1}^n) + \frac{\lambda^2 a^2}{2}(\bar{u}_{j+1}^n - 2\bar{u}_j^n + \bar{u}_{j-1}^n).$$

We can rearrange the terms to obtain

$$\bar{u}_j^{n+1} = \underbrace{\bar{u}_j^n - \lambda a(\bar{u}_j^n - \bar{u}_{j-1}^n)}_{\text{1st order upwind method}} - \underbrace{\frac{1}{2}a(1-\lambda a)(\bar{u}_{j+1}^n - 2\bar{u}_j^n + \bar{u}_{j-1}^n)}_{\text{“antidiffusive” scheme}}. \tag{73}$$

Focusing on the numerical fluxes, we can recast (73) as

$$F_{j+1/2}^{\mathrm{LW}} = a\bar{u}_j^n + \frac{a}{2}(1-\lambda a)(\bar{u}_{j+1}^n - \bar{u}_j^n). \tag{74}$$

In order to enforce the TVD property, the antidiffusive part of the numerical flux requires a limiting procedure:

$$F_{j+1/2}^{\mathrm{TVD}} = a\bar{u}_j^n + \frac{a}{2}(1-\lambda a)\Phi_{j+1/2}(\bar{u}_{j+1}^n - \bar{u}_j^n), \tag{75}$$

indicating that the limiter must be applied on the flux itself, in order to preserve the conservation form of the method:

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \lambda a(\bar{u}_j^n - \bar{u}_{j-1}^n) - \frac{1}{2}a(1-\lambda a)[\Phi_{j+1/2}(\bar{u}_{j+1}^n - \bar{u}_j^n) - \Phi_{j-1/2}(\bar{u}_j^n - \bar{u}_{j-1}^n)].$$

An alternative way of achieving high-order methods consists in concentrating on deriving a *more accurate reconstruction*. For instance, let us consider the following conservative scheme written in semidiscrete form (as (60))

$$\partial_t \bar{u}_j(t) = -\frac{1}{h}[F(v_j^+(t), v_{j+1}^-(t)) - F(v_{j-1}^+(t), v_j^-(t))]. \tag{76}$$

The values $v_j^{\pm}(t)$ are obtained by *linear* reconstruction from the cell averages $\bar{u}_j(t)$:

$$v_j^- = \bar{u}_j - \frac{1}{2}h\sigma_j, \qquad v_j^+ = \bar{u}_j + \frac{1}{2}h\sigma_j,$$

where the σ_j's are suitable slopes depending on the local cell averages $\bar{u}_j$.

This reconstructed (discrete) function v is therefore piecewise linear on the cells, but not necessarily continuous. A number of possibilities to select σ are available. An intuitive choice (already mentioned in the construction of Godunov schemes) is to take a *central slope*

$$\sigma_j = \frac{1}{2}[\frac{\bar{u}_{j+1} - \bar{u}_j}{h} + \frac{\bar{u}_j - \bar{u}_{j-1}}{h}] = \frac{\bar{u}_{j+1} - \bar{u}_{j-1}}{2h}.$$

In analogy with the central approximation of first order derivatives (which are $O(h^2)-$accurate), here we can apply a Taylor expansion for a sufficiently regular u to obtain

$$|v_j^-(t) - u(x_{j-1/2}, t)| = O(h^2), \qquad |v_j^+(t) - u(x_{j+1/2}, t)| = O(h^2).$$

Other options include one-sided slopes (of order $O(h)$):

$$\sigma_j = \begin{cases} \frac{\bar{u}_{j+1} - \bar{u}_j}{h} \\ \frac{\bar{u}_j - \bar{u}_{j-1}}{h} \end{cases} \text{ or },$$

and a class of methods called *slope-limiting schemes* consisting in forcing a cap for a given slope. This can be achieved, for example, using the minmod operator

$$\sigma_j = \mathrm{minmod}\left(\frac{\bar{u}_{j+1} - \bar{u}_j}{h}, \frac{\bar{u}_j - \bar{u}_{j-1}}{h}\right).$$