1. Lecture 1

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

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

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 problem. 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 discretization 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 principal of the shooting method for linear second order ODE boundary value problems (BVPs) of the form

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

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 $r(x)$:

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

and

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

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 exists. Rather than solving (1), one can replace the right boundary condition with a user specified slope at the left boundary

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

giving a parametrized solution, $y(x; s)$, for each s. It then remains to find a value of s, say s^*, such that its parametrization 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 $w \geq \frac{1}{2}h^2 Q^*$ 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

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

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

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

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 $\|y^p - y^q\|_\infty \leq \frac{\|y^1 - y^0\|}{1-\lambda}\lambda^{\min(p,q)}$ 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.

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 $(x_j + \theta h, y_k)$;

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

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

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

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.