Lecture 9: Numerical solution of boundary value problems

Initial vs. boundary value problems

In lectures 7 and 8 we discussed numerical solution techniques for initial value problems. Those concerned solutions of ordinary differential equations of the form

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(t, \mathbf{y}) \,, \tag{1}$$

where **initial** conditions were imposed at the same locations, most likely $t = 0$ in time, of the form

$$\mathbf{y}(0) = \mathbf{y}_0 \,. \tag{2}$$

That is, every initial value of the elements of $\mathbf{y}$ is specified at the same location in time.

An example of an initial value problem is given by the second order ODE

$$\frac{d^2y}{dt^2} + g = 0 \,, \tag{3}$$

with initial conditions $y(0) = y_0$ and $\dot{y}(0) = 0$. This is written in vector form as

$$\frac{d\mathbf{y}}{dt} + \mathbf{f}(t, \mathbf{y}) = 0 \,, \tag{4}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y \\ \dot{y} \end{pmatrix} \,, \tag{5}$$

and

$$\mathbf{f}(t, \mathbf{y}) = \begin{pmatrix} -y_2 \\ g \end{pmatrix} \,, \tag{6}$$

with initial conditions

$$\mathbf{y}(0) = \begin{pmatrix} y_0 \\ 0 \end{pmatrix} \,. \tag{7}$$

The difference between initial and boundary value problems is that rather than initial conditions being imposed at the same point in the independent variable (in this case, t), boundary conditions are imposed at different values of the independent variable. As an example of a boundary value problem, consider the second order ODE

$$\frac{d^2y}{dx^2} + \lambda^2 y = 0 \,, \tag{8}$$

with boundary conditions given by $y(0) = 0$ and $y(1) = 1$. This problem cannot be solved using the methods we learned for the initial value problems because the two conditions imposed on the problem are not at coincident locations of the independent variable x.

Boundary condition types

Dirichlet condition (Value specified)

When the value is specified at a particular location of the independent variable, this is known as a Dirichlet boundary condition. Examples of a Dirichlet boundary condition are given by

$$y(0) = a\,, \tag{9}$$

or

$$y(b) = 2\,. \tag{10}$$

Neumann condition (Derivative specified)

If the derivative is specified, then this is known as a Neumann boundary condition. Examples of Neumann conditions are given by

$$y'(0) = 1\,, \tag{11}$$

and

$$y'(a) = b\,. \tag{12}$$

Mixed condition (Gradient + value)

When the boundary condition specifies an equation that involves both a value and the derivative, it is known as a mixed condition. Examples are given by

$$y'(a) + \lambda y(a) = 0\,, \tag{13}$$

and

$$y'(0) = 2y(0)\,. \tag{14}$$

The shooting method

The shooting method uses the methods developed for solving initial value problems to solve boundary value problems. The idea is to write the boundary value problem in vector form and begin the solution at one end of the boundary value problem, and "shoot" to the other end with an initial value solver until the boundary condition at the other end converges to its correct value.

The vector form of the boundary value problem is written in the same way as it was for the initial value problems, except all of the initial conditions are not known a-priori. As an example, take the boundary value problem

$$\frac{d^2y}{dx^2} + \lambda^2 y = 0\,, \tag{15}$$

with boundary conditions $y(0) = 0$ and $y(1) = 1$. In vector form, this is given by

$$\frac{d\mathbf{y}}{dx} + \mathbf{f}(x, \mathbf{y}) = 0\,, \tag{16}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y \\ y' \end{pmatrix} , \tag{17}$$

and

$$\mathbf{f}(x, \mathbf{y}) = \begin{pmatrix} -y_2 \\ \lambda^2 y_1 \end{pmatrix} . \tag{18}$$

All of the elements of the boundary condition vectors are not known initially, because certain components will depend on the solution of the problem. Since we are only given $y(0)$ and $y(1)$, then the boundary condition vectors are given by

$$\mathbf{y}(0) = \begin{pmatrix} 0 \\ ? \end{pmatrix} , \mathbf{y}(1) = \begin{pmatrix} 1 \\ ? \end{pmatrix} . \tag{19}$$

We leave question marks in place of the unknown boundary conditions because they will only be known when we actually solve the problem. In this case, we will only know the values of $y'(0)$ and $y'(1)$ when we have the solution to the boundary value problem (15).

As another example, suppose we want to express the boundary value problem

$$y_{xxxx} + a y_{xx} = 0 , \tag{20}$$

with boundary conditions $y(0) = 0$, $y'(0) = 1$, $y(1) = 0$, and $y'(0) = -1$ in vector form. Because this is a fourth order ODE, we know that it has four elements in the $\mathbf{y}$ vector, and as a result, it has the four given boundary conditions. The $\mathbf{y}$ vector is given by

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} y \\ y_x \\ y_{xx} \\ y_{xxx} \end{pmatrix} , \tag{21}$$

and the boundary value problem is given by

$$\frac{d\mathbf{y}}{dx} + \mathbf{f}(x, \mathbf{y}) = 0 , \tag{22}$$

where

$$\mathbf{f}(x, \mathbf{y}) = \begin{pmatrix} -y_2 \\ -y_3 \\ -y_4 \\ a y_3 \end{pmatrix} , \tag{23}$$

with boundary conditions

$$\mathbf{f}(0) = \begin{pmatrix} 0 \\ 1 \\ ? \\ ? \end{pmatrix} , \mathbf{f}(1) = \begin{pmatrix} 0 \\ -1 \\ ? \\ ? \end{pmatrix} . \tag{24}$$

Because we are only given four boundary conditions, the other values of the derivatives at the boundary are determined after a solution of the problem is found.

The best way to illustrate the shooting method is with an example.

An example of the shooting method

Find the solution of the boundary value problem

$$\frac{d^2y}{dx^2} - y = 0\,, \tag{25}$$

with boundary conditions $y(0) = 0$, $y'(1) = -1$.

1: Write the BVP in vector form

In order to solve this problem numerically, we write it in its vector form as

$$\frac{d\mathbf{y}}{dx} + \mathbf{f}(x, \mathbf{y}) = 0\,, \tag{26}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y \\ y_x \end{pmatrix}\,, \tag{27}$$

and

$$\mathbf{f}(x, \mathbf{y}) = \begin{pmatrix} -y_2 \\ -y_1 \end{pmatrix}\,, \tag{28}$$

with boundary conditions

$$\mathbf{y}(0) = \begin{pmatrix} 0 \\ ? \end{pmatrix}\,, \mathbf{y}(1) = \begin{pmatrix} ? \\ -1 \end{pmatrix}\,. \tag{29}$$

2: Discretize

The problem is first discretized into N points, the number of which depends on the desired accuracy of the solution. We will use $N = 20$ for this example and assume that this yields a converged result. The independent variable x is discretized with $x_i = i\Delta x$, with $\Delta x = L/(N-1)$, where $L = 1$ is the size of the domain. Sometimes we might need to discretize the grid with an unequispaced grid if the terms in the boundary value problem vary considerably in some locations of the domain in which we are solving the problem. Since this problem is linear and behaves smoothly, we do not need to worry about this.

3: Choose an integrator

For this problem we will use the Euler predictor-corrector algorithm, which will give us values for y_1 and y_2 in the domain if we give it starting values $y_1(0)$ and $y_2(0)$. But only $y_1(0)$ is specified, so we need to iterate to determine $y_2(0)$.

4: Iterate to find the solution

This is the trickiest part of the problem. Because the only boundary condition at $x = 0$ is $y_1(0) = 0$, then we need to guess the value of $y_2(0)$ and use the predictor-corrector algorithm to shoot to the other end of the domain and see if this guess satisfies the boundary condition

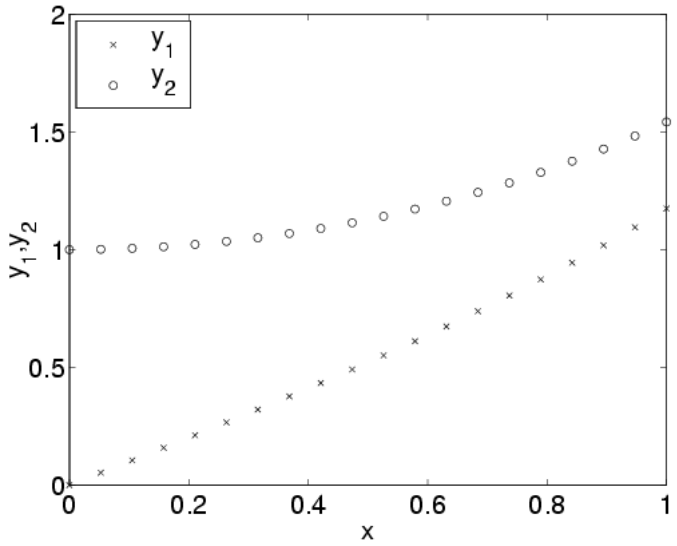

Figure 1: Results of the shooting method with a guess of $y_2(0) = 1$ which yields $y_2(1) = 1.542566$.

$y_2(1) = -1$. Let's say we guess a value of $y_2(0) = 1$. The predictor-corrector algorithm will yield the result shown in Figure 1. Because $y_2(1) = 1.542566$ does not match the correct value of $y_2(1) = -1$ (which is specified as a boundary condition), then we need to try again. Let's try $y_2(0) = -1$. Using this guess, the predictor-corrector shooting method yields the result shown in Figure 2. Again, this is the incorrect answer since a guess of $y_2(0) = -1$

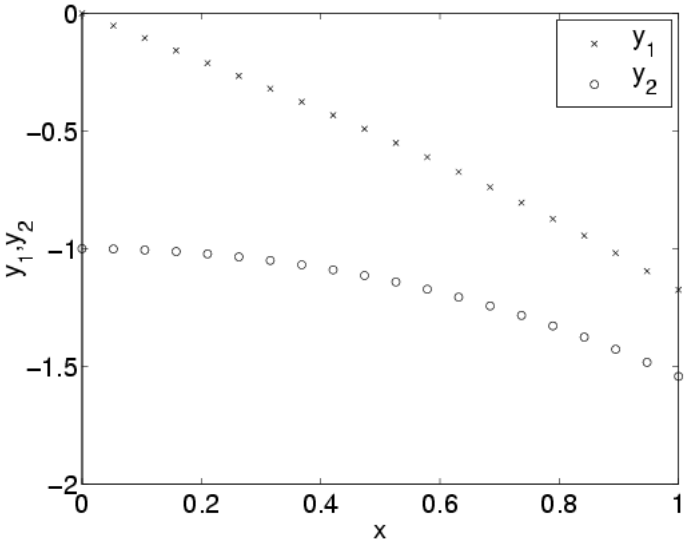

Figure 2: Results of the shooting method with a guess of $y_2(0) = -1$ which yields $y_2(1) = -1.542566$.

yields $y_2(1) = -1.542566$.

The shooting method gives us a value for $y_2(1)$ when we are given a value for $y_2(0)$. That is, if we guess the slope y_x at $x = 0$, then the shooting method will give us a value for the

slope y_x at $x = 1$, which is specified as a boundary condition in the problem as $y'(1) = -1$. To solve the boundary value problem, we need to iterate with different values of $y_2(0)$ until we converge upon the correct value of $y_2(1)$. This can be done with a root-finder such as the bisection method, the secant method, or linear interpolation. The table below depicts the results of the two previous guesses we used to solve the initial value problem. We can use

Guess number	Guess for $y_2(0)$	Result of shooting method $y_2(1)$
1	1.0	1.542566
2	-1.0	-1.542566

the Secant method to find a good value for the next guess. If we let s be the guess for $y_2(0)$ and $E(s) = y_2(1) - y'(1)$ be the error in the result of the shooting method, then we need to use the Secant method to find the root of

$$E(s) = 0\,. \tag{30}$$

This is done by using the formula for the secant method, which is given by

$$s_3 = s_2 - E(s_2)\left(\frac{s_1 - s_2}{E(s_1) - E(s_2)}\right)\,. \tag{31}$$

Using the results from the table above, we have

$$\begin{aligned} E(s_1) &= 1.542566 - (-1) = 2.542566\,, \\ E(s_2) &= -1.542566 - (-1) = -0.542566\,, \end{aligned}$$

and

$$s_3 = -1.0 - (-0.542566)\left(\frac{1.0 - (-1.0)}{2.542566 - (-0.542566)}\right) = -0.648270\,. \tag{32}$$

If we use $y_2(0) = -0.648270$, then the result is shown in Figure 3. As shown in the figure, when we use a gues of $y_2(0) = -0.64827$, we end up with a slope at $x = 1$ of $y_2(1) = -0.999999$, which is the exact value (or close enough)! The result in Figure 3 is therefore the solution of the boundary value problem, which is $y = -\sinh(x)/\cosh(1)$. From this we can see that the shooting method only requires us to shoot for the result three times for **linear** boundary value problems. Two guesses are required, and then a linear interpolation yields the solution to within the errors of the method used to integrate the ODE. In this case, since the Euler predictor-corrector method is second-order accurate in Δx, then we know that we must have the solution to the boundary value problem to within $\mathcal{O}\left(\Delta x^2\right)$.

Only three steps are required to find the solution for linear problems, and the accuracy of the result is governed by the accuracy of the shooting method used. For nonlinear problems, however, more iterations are required, and one must continue to integrate until the residual error in the root of $E(s)$ is below some specified tolerance. If the tolerance is less than the error of the shooting method, then the error in the solution of the boundary value problem will be governed by the shooting method.

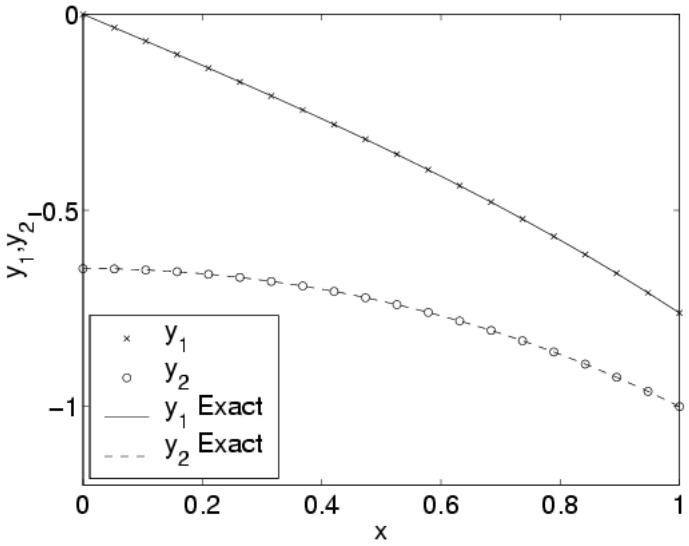

Figure 3: Results of the shooting method with a guess of $y_2(0) = -0.648270$ which yields $y_2(1) = -0.999999$.

The finite-difference method

The boundary value problem is given by

$$\frac{d^2y}{dx^2} - y = 0\,, \tag{33}$$

with boundary conditions $y(0) = 0$, $y'(1) = -1$. In order to solve this boundary value problem with the finite difference method, the following steps should be taken.

1: Discretize x

The discretization of the boundary value problem for the finite-difference method is done differently than for the shooting method. In order to guarantee second order accuracy of the Neumann (derivative) boundary condition at $x = 1$ (and lead to a tridiagonal system as in step 5), the grid must be **staggered** about that boundary. That is, the x values must lie on either side of the point $x = 1$. In order to stagger the grid, a discretization of x with N points must be given by

$$x_i = \left(i - \frac{3}{2}\right)\Delta x\,, \quad \text{(Neumann boundary conditions)} \tag{34}$$

with $\Delta x = 1/(N-2)$. This is the discretization we will use for the current problem, since it has a Neumann boundary condition.

As an aside, if the problem only consists of Dirichlet boundary conditions, then it is better to collocate the x values with the boundaries. In this case it is best to use the discretization

$$x_i = i\Delta x\,, \quad \text{(Dirichlet boundary conditions)} \tag{35}$$

with $\Delta x = 1/(N-1)$.

2: Discretize the governing ODE

The governing ODE for this problem can be discretized by rewriting it as a finite difference equation at each point x_i, for which

$$\left.\frac{d^2y}{dx^2}\right|_i - y_i = 0 \quad i = \{2, \ldots, N-1\}\,. \tag{36}$$

The second order accurate finite difference approximation is then given by

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{\Delta x^2} - y_i = 0\,, \tag{37}$$

which can be rewritten as

$$a_i y_{i-1} + b_i y_i + c_i y_{i+1} = d_i\,, \tag{38}$$

where

$$\begin{aligned} a_i &= \frac{1}{\Delta x^2}\,, \\ b_i &= -\left(1 + \frac{2}{\Delta x^2}\right)\,, \\ c_i &= \frac{1}{\Delta x^2}\,, \\ d_i &= 0\,. \end{aligned}$$

We have neglected the discretization error, keeping in mind that the discretization is second order accurate in Δx, and we will assume that $d_i \neq 0$ and that the coefficients are not constant with i to be as general as possible. These equations are only valid for $i \in \{2, \ldots, N-1\}$ since the discrete second derivative is not defined at $i = 1$ or $i = N$ as we have written it.

3: Discretize the boundary conditions

Just as the governing ODE is discretized, so must the boundary conditions. The boundary condition at $x = 0$ is given by $y(0) = 0$. Because the grid we are using is staggered, we do not have values at $x = 0$, but rather, we have values at $x_1 = -\Delta x/2$ and $x_2 = +\Delta x/2$. Therefore, the value at $x = 0$ must be interpolated with the values at y_1 and y_2. This is given by a centered interpolation to obtain $y_{3/2}$ as

$$y_{3/2} = \frac{y_1 + y_2}{2} + \mathcal{O}\left(\Delta x^2\right) = 0\,. \tag{39}$$

Solving for y_1 and neglecting the discretization error, we have

$$y_1 = -y_2\,. \tag{40}$$

The boundary condition at $x = 1$ is discretized by writing the second-order accurate approximation for the first derivative at $x = 1$ to obtain

$$\left.\frac{dy}{dx}\right|_{i=N-1/2} = \frac{y_N - y_{N-1}}{\Delta x} + \mathcal{O}\left(\Delta x^2\right) = -1\,. \tag{41}$$

Leaving out the discretization error, we have

$$y_N = y_{N-1} - \Delta x\,. \tag{42}$$

4: Embed the boundary conditions

The discretized ODE (38) is only valid for $i \in \{2, \ldots, N-1\}$. Therefore, it can only be used to solve for points in that range. Any terms in the discretized ODE that contain points not in that range are removed by embedding the boundary conditions. If we write the discretized ODE at $i = 2$ and $i = N-2$ we have

$$\begin{array}{ccccccccc} a_2 y_1 & + & b_2 y_2 & + & c_2 y_3 & = & d_2\,, \\ a_{N-1} y_{N-2} & + & b_{N-1} y_{N-1} & + & c_{N-1} y_{N-1} & = & d_{N-1}\,. \end{array} \tag{43}$$

From the boundary conditions, we know that

$$\begin{aligned} y_1 &= -y_2\,, \\ y_N &= y_{N-1} - \Delta x\,. \end{aligned}$$

Substituting the boundary conditions into equations (43), we have

$$\begin{array}{ccccccc} & & (b_2 - a_2) y_2 & + & c_2 y_3 & = & d_2\,, \\ a_{N-1} y_{N-2} & + & (b_{N-1} - c_{N-1}) y_{N-1} & & & = & d_{N-1} + c_{N-1}\Delta x\,. \end{array} \tag{44}$$

5: Set up the linear system

The discretized set of ODEs that govern the behavior of y_i where $i \in \{2, \ldots, N-1\}$ is then given by

$$\begin{array}{ccccccccccc} i = 2 & & (b_2 - a_2) y_2 & + & c_2 y_3 & & & & = & & d_2\,, \\ i = \{3, \ldots N-2\} & & a_i y_{i-i} & + & b_i y_i & + & c_i y_{i+1} & & = & & d_i \\ i = N-1 & & & & a_{N-1} y_{N-2} & + & (b_{N-1} - c_{N-1}) y_{N-1} & & = & & d_{N-1} + c_{N-1}\Delta x\,. \end{array}$$

This represents a linear system of the form

$$\begin{pmatrix} b_2 & c_2 & & & & & \\ a_3 & b_3 & c_3 & & & & \\ & a_4 & b_4 & c_4 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & a_{N-3} & b_{N-3} & c_{N-3} & \\ & & & & a_{N-2} & b_{N-2} & c_{N-2} \\ & & & & & a_{N-1} & b_{N-1} \end{pmatrix} \begin{pmatrix} y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_{N-3} \\ y_{N-2} \\ y_{N-1} \end{pmatrix} = \begin{pmatrix} d_2 \\ d_3 \\ d_4 \\ \vdots \\ d_{N-3} \\ d_{N-2} \\ d_{N-1} \end{pmatrix},$$

where we have performed the replacements

$$\begin{aligned} b_2 &\leftarrow b_2 - a_2\,, \\ b_{N-1} &\leftarrow b_{N-1} + c_{N-1}\,, \\ d_{N-1} &\leftarrow d_{N-1} + c_{N-1}\Delta x\,. \end{aligned} \tag{45}$$

6: Solve the linear system

The linear system derived in the previous step can be represented as

$$\mathbf{A}\mathbf{y} = \mathbf{d}\,. \tag{46}$$

The objective is to now solve the system with

$$\mathbf{y} = \mathbf{A}^{-1}\mathbf{d}\,. \tag{47}$$

We can usually take advantage of the structure of $\mathbf{A}$ in order to speed up the calculation of its inverse. In this case it turns out that $\mathbf{A}$ is a tridiagonal matrix. That is, it has three diagonals, and as a result, it can be solved with the use of a tridiagonal solver.

The solution $\mathbf{y}$ then represents the solution of the boundary value problem we initially set out to solve. Due to the accumulation of errors in the tridiagonal solver, this method turns out to be first-order accurate in Δx, as opposed to the second-order accurate shooting method with the use of the Euler predictor-corrector method.