IMT2112 - The finite difference method

Elwin van 't Wout*

August 31, 2024

Abstract

Many physical phenomena can be modelled with a boundary value problem. In special cases, these systems of partial differential equations and boundary conditions have analytical solutions but for most problems of engineering interest the solution has to be approximated numerically. A common algorithm to numerically solve boundary value problems is the finite difference method. This short report explains the basics of this numerical method and applies it to the Helmholtz equation for harmonic wave propagation.

1 The Helmholtz equation

Wave propagation problems are governed by the wave equation

$$\frac{\partial^2 u}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = 0 \tag{1}$$

where $u = u(x,t)$ denotes the unknown wave field at location x and time t, and c a constant wavespeed. The wave equation allows for separable solutions, that is, $u(x,t) = v(x)w(t)$. Since many problems of engineering interest have incident wave fields that are oscillating in time with a fixed frequency, we will consider a harmonic wave of the form $w(x) = \cos(\omega t)$ with ω the angular frequency. Then, the spatial unknown function $v(x)$ has to satisfy the Helmholtz equation

$$-\frac{d^2 v}{dx^2} - k^2 v = 0 \tag{2}$$

*Instituto de Ingeniería Matemática y Computacional, Pontificia Universidad Católica de Chile, e.wout@uc.cl

where $k = \omega/c$ denotes the wavenumber. Here, we will assume $v(x)$ and k to be real-valued. However, in many applications, complex-valued functions need to be used, where the imaginary part of the wavenumber models damping.

Exercise 1. Show that if $u(x,t) = v(x)\cos(\omega t)$ satisfies the wave equation, then $v(x)$ needs to satisfy the Helmholtz equation.

In the general case of the entire one-dimensional domain, i.e., $x \in \mathbb{R}$, the general solution of the Helmholtz equation is given by $v(x) = C_1\cos(kx) + C_2\sin(kx)$ for arbitrary constants C_1 and C_2. However, when domains with boundaries are considered, analytical solutions are only available in rare cases. In other words, analytical solutions of the Helmholtz equation are difficult to obtain. This becomes even more complicated when the boundary conditions are given at three-dimensional structures. Hence, in most realistic cases we need to use numerical methods that approximate the solution of the Helmholtz equation.

Example 1. Consider the boundary value problem

$$\begin{cases} -\frac{d^2v}{dx^2} - k^2 v = 0, \quad 0 < x < \pi; \\ v(0) = 0; \\ v(\pi) = 0. \end{cases} \tag{3}$$

One possible solution would be $v(x) = \sin(kx)$, which satisfies the Helmholtz equation and $v(0) = 0$. However, $v(\pi) = \sin(k\pi) = 0$ if and only if k is an integer. Therefore, this solution only satisfies the Helmholtz system for integer k. For non-integer k, the only explicit solution is $v(x) = 0$.

Remark 1. In several cases, solutions in the form of infinite series can be obtained by considering the Fourier transform of the Helmholtz equation.

Remark 2. In general, the Helmholtz equation has a non-zero right-hand side: $-\frac{d^2v}{dx^2} - k^2 v = f(x)$ for a given function f. This function typically represents sources or sinks of energy. Since the solution will depend on f, the feasibility of finding an analytical solution also depends on the definition of f. Furthermore, solutions of the differential equation with zero right-hand side are often related to resonances.

2 Numerical discretisation

In most cases, no analytical solution of *boundary value problems* (a differential equation with boundary conditions) is available, even if existence and

uniquess of the solution can be proven. Hence, numerical methods need to be used to approximate the solution. Most numerical methods follow the approach of *discretisation*, that is, instead of looking for an explicit function that satisfies the differential equation, the solution is approximated on a finite set of points or intervals.
The numerical discretisation brings two fundamental questions:

1. how to obtain an approximation of the solution on the points, and
2. how many points do we need to get an accurate approximation.

Let us first look into the first question. In fact, many different approximation techniques exist, of which the most common ones are *finite differences*, *finite volumes*, and *finite elements*. Other discretisation methods, with a different approach, are *boundary element* and *spectral methods*. Here, we will focus on the finite difference method (ES: *método de las diferencias finitas*). Answering the second question depends on the numerical method used and the equation to be solved. More information on this for wave problems will be given in Section 5.

3 The finite difference method

Let us consider the Helmholtz system

$$\begin{cases} -\frac{d^2v}{dx^2} - k^2 v = f, \quad 0 < x < 1; \\ v(0) = 0; \\ v(1) = 0 \end{cases} \tag{4}$$

where $f = f(x)$ is a known function and $v = v(x)$ the unkown wave field. The function f typically models the source of the wave field and, when non-zero, prevents the trivial solution of $v(x) = 0$ to be valid. There is no obvious analytical solution for general k and f and we, therefore, need an approximation method.

3.1 Computational grid

The first step for numerical methods is to define a *grid* or *mesh* (ES: *malla*), that is, the discrete points in which to approximate the solution. Let us consider an equidistant grid, where the points are uniformly distributed over the interval. That is,

$$x_i = ih = i\frac{1}{N}, \text{ for } i = 0, 1, 2, \ldots, N \tag{5}$$

where h is called the *mesh width* or *grid spacing* (ES: *paso* or *ancho de subintervalo*) and each x_i a *node* (ES: *nodo*).

3.2 Finite difference approximation

Remember that the goal is to find a good approximation of the solution of the boundary value problem in each and every point of the computational grid. Hence, let us consider the boundary nodes first and then the interior nodes.
The approximation in nodes x_0 and x_N is straightforward, namely the boundary conditions $v(x_0) = v(0) = 0$ and $v(x_N) = v(1) = 0$, which are called Dirichlet conditions. Other boundary conditions can be used as well, for example Neumann conditions that depend on the derivative, such as $v'(0) = 0$. Notice that for Neumann boundary conditions, the value of v on the boundary is unknown and has to be approximated as well.
For the interior nodes, the only condition is that the function needs to satisfy the Helmholtz equation, that is,

$$-\frac{d^2v}{dx^2}(x_i) - k^2 v(x_i) = f(x_i), \text{ for } i = 1, 2, \ldots, N-1. \tag{6}$$

Since we need to find approximations of $v(x)$ in the nodes x_i only, let us consider the set of unknowns v_i that approximate the solution $v(x_i)$. The aim is to find conditions for these unknowns. Hence, we need to write each $\frac{d^2v}{dx^2}(x_i)$ as a function of v_j for $j = 1, 2, \ldots, N-1$. Notice that we can couple the approximation in one node with the approximation in neighbouring nodes. This leads us to the idea of using a Taylor series, that is, for sufficiently smooth $v(x)$, we have

$$v(x_{i+1}) = v(x_i) + h\frac{dv}{dx}(x_i) + \frac{1}{2}h^2\frac{d^2v}{dx^2}(x_i) + \frac{1}{6}h^3\frac{d^3v}{dx^3}(x_i) + \mathcal{O}(h^4), \tag{7}$$

$$v(x_{i-1}) = v(x_i) - h\frac{dv}{dx}(x_i) + \frac{1}{2}h^2\frac{d^2v}{dx^2}(x_i) - \frac{1}{6}h^3\frac{d^3v}{dx^3}(x_i) + \mathcal{O}(h^4). \tag{8}$$

In this context, the big-O notation of $f(x) = \mathcal{O}(h^n)$ means that there exists a constant M such that $|f(x)| \leq M|h|^n$ for all $|x - x_i| \leq h$. Taking the difference of the two Taylor series leads to

$$\frac{v(x_{i+1}) - v(x_{i-1})}{2h} = \frac{dv}{dx}(x_i) + \mathcal{O}(h^2), \tag{9}$$

which means that $(v(x_{i+1}) - v(x_{i-1}))/(2h)$ is a good approximation of $\frac{dv}{dx}(x_i)$ for small h. This is called a *central finite difference approximation* of the

first derivative. For the second derivative, we need another formula, that is,

$$\frac{v(x_{i+1}) - 2v(x_i) + v(x_{i-1})}{h^2} = \frac{d^2v}{dx^2}(x_i) + \mathcal{O}(h^2). \tag{10}$$

The $\mathcal{O}(h^2)$ term is called the *truncation error* (ES: *eror de truncamiento*).

Exercise 2. When using the grid (5) for a finite difference approximation of the Helmholtz system (4), how many unknowns do we have? In other words, in how many nodes do we need to approximate the solution?

Exercise 3. Show that the finite difference formula (10) has a truncation error of $O(h^2)$ and not just $O(h)$.

Exercise 4. Show that the forward finite difference formula

$$\frac{dv}{dx}(x_i) \approx \frac{v(x_{i+1}) - v(x_i)}{h}$$

has a truncation error of $O(h)$.

3.3 Finite difference matrix

Given the finite difference approximation (10) of the second derivative of the solution, we can use the following relation:

$$\frac{d^2v}{dx^2}(x_i) \approx \frac{v_{i+1} - 2v_i + v_{i-1}}{h^2}. \tag{11}$$

Substitution of this into the Helmholtz equation results in

$$\frac{-v_{i+1} + 2v_i - v_{i-1}}{h^2} - k^2 v_i \approx f(x_i) \tag{12}$$

for every interior node of the mesh.
In order to obtain actual values of v_i we will solve the set of equations

$$\frac{-v_{i+1} + 2v_i - v_{i-1}}{h^2} - k^2 v_i = f_i \text{ for } i = 1, 2, \ldots, N-1 \tag{13}$$

because the finite difference approximations have small errors when there are sufficient nodes in the mesh. The boundary conditions state $v_0 = 0$ and $v_N = 0$. Notice that we now have $N-1$ linear equations for $N-1$ unknowns. In a matrix-vector notation, this would be

$$A\mathbf{u} = \mathbf{f}, \tag{14}$$

where

$$A = \frac{1}{h^2}\begin{bmatrix} 2-(kh)^2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2-(kh)^2 & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 2-(kh)^2 & -1 \\ 0 & \dots & \dots & 0 & -1 & 2-(kh)^2 \end{bmatrix}$$

is the discretisation matrix, $\mathbf{u} = [u_1, u_2, \dots, u_{N-1}]^T$ the vector of unknowns, and $\mathbf{f} = [f(x_1), f(x_2), \dots, f(x_{N-1})]^T$ the known right-hand-side vector. Now, solving the set of linear equations gives the vector $\mathbf{u}$ and, therefore, an approximation of the solution in each node of the grid.

Exercise 5. Show that if $k = 0$, we have

$$\langle \mathbf{x}, h^2 A\mathbf{x} \rangle = x_1^2 + \sum_{m=1}^{N-2} (x_{m+1} - x_m)^2 + x_{N-1}^2$$

where the brackets denote the vector dot product. Explain from this result that the matrix A is positive definite for $k = 0$.

Exercise 6. Use the circle theorem of Gershgorin to show that the matrix A is positive definite if $\Re(k^2) < 0$. (Remember that k can be complex-valued.)

Exercise 7. Derive a finite difference approximation and the resulting discretisation matrix for the convection-diffusion problem given by

$$\begin{cases} -a\frac{d^2v}{dx^2} + b\frac{dv}{dx} = f(x), \quad 0 < x < 1; \\ v(0) = 0; \\ v(1) = 1 \end{cases} \tag{15}$$

for positive parameters a and b.

Remark 3. For wave problems, the wavenumber k is usually real and positive, leading to an indefinite A. In general, indefinite systems have a large condition number and are, therefore, difficult to solve numerically. This is an important topic in computational sciences and the area of *numerical linear algebra* deals with the development of fast solvers for linear equations.

Remark 4. Solutions of the homogeneous Helmholtz equation $-v'' - k^2 v = 0$ that satisfy the boundary conditions are called *resonance modes* and only appear at specific values of k. These nonzero solutions are in the nullspace of the Helmholtz operator and therefore result in singularities of the discretisation matrix as well.

Remark 5. Many authors prefer to write the Helmholtz equation as $-v'' - k^2 v = 0$ instead of the equivalent form of $v'' + k^2 v = 0$. The finite difference matrix of the Laplace operator has negative values on the main diagonal, whereas the the matrix for the negative Laplace operator has positive values. Also, the finite difference matrix of $-v''$ is positive definite, instead of negative definite for v''. Since it is more intuitive to use positive definite matrices, $-v''$ is typically used in Helmholtz equations.

3.4 Neumann boundary conditions

Let us consider the Helmholtz system

$$\begin{cases} -\frac{d^2v}{dx^2} - k^2 v = f, \quad 0 < x < 1; \\ v(0) = 0; \\ \frac{dv}{dx}(1) = 0 \end{cases} \tag{16}$$

where at the right boundary a Neumann condition is used.

Exercise 8. Show that $v(x) = \sin(kx)$ is a solution of this boundary value problem for $f = 0$ and specific values of k.

The finite difference method explained in Section 3.2 will not change for the left boundary condition and the Helmholtz equation. For the right boundary condition, we do not know the value of $v(x_N)$ so we cannot do the same as before, and we need to treat u_N as an additional unknown. Notice that we can use the finite difference approximation

$$\frac{dv}{dx}(x_N) \approx \frac{u_N - u_{N-1}}{h} \tag{17}$$

which is first-order accurate. The resulting discretisation matrix is then given by

$$A = \frac{1}{h^2} \begin{bmatrix} 2-(kh)^2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2-(kh)^2 & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 2-(kh)^2 & -1 \\ 0 & \dots & \dots & 0 & -h & h \end{bmatrix}$$

for the solution vector $\mathbf{u} = [u_1, u_2, \ldots, u_{N-1}, u_N]^T$.

Exercise 9. Show that if the truncation error of the Neumann boundary condition is $\mathcal{O}(h)$ and for all other unknowns $\mathcal{O}(h^2)$, the overall error is $\mathcal{O}(h)$.

Exercise 10. Explain why the approximation $\frac{dv}{dx}(x_N) \approx \frac{u_{N+1}-u_{N-1}}{h}$, which is second-order accurate, cannot be used for the Neumann boundary condition.

Exercise 11. Derive a finite-difference approximation for the Neumann boundary condition that is second-order accurate.
Hint: use the Taylor polynomial for $u(x_{N-2})$ with respect to x_N.

3.5 Two dimensional problems

Let us consider the two-dimensional Helmholtz system

$$\begin{cases} -\frac{d^2v}{dx^2} - \frac{d^2v}{dy^2} - k^2 v = f, & 0 < x, y < 1; \\ v = 0, & \text{for } x = 0,\ x = 1,\ y = 0,\ \text{and } y = 1 \end{cases} \tag{18}$$

for the unknown $u = u(x, y)$.

Exercise 12. Show that the function $u(x, y) = \sin(k_x x)\sin(k_y y)$ is a solution of the homogeneous Helmholtz equation $-\nabla^2 u - k^2 u = 0$, but only when k_x and k_y satisfy a specific condition.

In most cases, this Helmholtz problem does not have an explicit solution, so we need numerical methods to approximate the solution. In order to use the finite difference method, we first need to define the grid, which is now given by a set of points in a rectangle. Let us use an equidistant grid given by

$$\mathbf{x}_{i,j} = \begin{bmatrix} x_i \\ y_j \end{bmatrix} = \begin{bmatrix} ih_x \\ jh_y \end{bmatrix} \text{ for } i = 0, 1, 2, \ldots, N_x \text{ and } j = 0, 1, 2, \ldots, N_y \tag{19}$$

where $h_x = 1/N_x$ and $h_y = 1/N_y$ the mesh sizes. The finite difference approximation for interior nodes is now given by

$$\frac{-u_{i-1,j} + 2u_{i,j} - u_{i+1,j}}{h_x^2} + \frac{-u_{i,j-1} + 2u_{i,j} - u_{i,j+1}}{h_y^2} - k^2 u_{i,j} = f_{i,j}. \tag{20}$$

for the unknowns $u_{i,j}$.

Exercise 13. How many unknowns are there in the finite difference method? In other words, in how many nodes do we need to approximate the solution?

Exercise 14. Show that this finite difference approximation is second-order accurate.

If we want to write two-dimensional finite differences in a matrix-vector form $A\mathbf{u} = \mathbf{b}$, we need to number the nodes in a one-dimensional way. The easiest approach is to count the nodes row-by-row or column-by-column. These are called a horizontal and vertical numbering, respectively. Both approaches are called *lexicographical* numbering. The numbering can be chosen freely, and many other numbering schemes exist (diagonal, spiral, etc.).

Exercise 15. For a horizontal numbering, give the conversion formula from the double index (i, j) to a single index n for the interior nodes.

Exercise 16. For $N_x = 4$ and $N_y = 3$, derive the discretisation matrix of the Helmholtz problem (18). How does the matrix change between a horizontal and vertical numbering?

4 Sparse linear algebra

The finite difference methods use a local approximation for the solutions of boundary value problems. This locality can also be observed in the linear system to solve. The discretisation matrix (14) has nonzero elements on the three main diagonals only. Matrices that mostly consist of zeros are called *sparse* (ES: *rala*). The linear algebra, such as the matrix-vector multiplication can be implemented more efficiently for sparse matrices: the zeros do not have to be stored.

Exercise 17. How many nonzero elements does the discretisation matrix (14) have? Specify your answer in terms of N.

Exercise 18. How many nonzero elements does the discretisation matrix (14) have in the case of a two-dimensional Helmholtz equation (18) with central finite differences?
Hint: the matrix is called *pentadiagonal* for lexicographical ordering.

Remark 6. There is no clear definition of a *sparse* matrix. A common guideline is that matrices $A \in \mathbb{R}^{n\times n}$ that have $\mathcal{O}(n)$ nonzero elements are sparse. Another guideline is that matrices are called sparse when storing the matrix in a specialised sparse format is more efficient than dense storage.

4.1 Stencil notation

Finite difference approximations have typically the same structure for all different nodes, where relations are specified between neighbouring nodes. The stencil notation is a convenient way of representing finite difference approximations. Notice that equation (20) can be written as

$$\begin{bmatrix} 0 & -\frac{1}{h_y^2} & 0 \\ -\frac{1}{h_x^2} & \frac{2}{h_x^2}+\frac{2}{h_y^2}-k^2 & -\frac{1}{h_x^2} \\ 0 & -\frac{1}{h_y^2} & 0 \end{bmatrix} \star \begin{bmatrix} u_{i-1,j+1} & u_{i,j+1} & u_{i+1,j+1} \\ u_{i-1,j} & u_{i,j} & u_{i+1,j} \\ u_{i-1,j-1} & u_{i,j-1} & u_{i+1,j-1} \end{bmatrix} = f_{ij} \qquad (21)$$

where the $\star$ denotes a *reduction*, that is, a pointwise multiplication of the matrices followed by the summation of each element. The first matrix is called the *stencil* of the finite difference approximation and is often presented as

$$\begin{bmatrix} & -\frac{1}{h_y^2} & \\ -\frac{1}{h_x^2} & \frac{2}{h_x^2}+\frac{2}{h_y^2}-k^2 & -\frac{1}{h_x^2} \\ & -\frac{1}{h_y^2} & \end{bmatrix}$$

where the empty elements are zero. Notice that the stencil is not considered to be an actual matrix, it is more a convenient notation of describing a finite difference approximation.
More general, the stencil corresponding to node (i, j) can be written as

$$\begin{bmatrix} & N_{ij} & \\ W_{ij} & C_{ij} & E_{ij} \\ & S_{ij} & \end{bmatrix}$$

where the values correspond to the coefficient for the north, east, south and west neighbour as well as the center contribution. In the specific finite difference approximation (20), all elements of the stencil are constant but, in general, the elements of the stencil depend on the location in the mesh.

Remark 7. Stencils can have different shapes, for example with five elements in a cross or with nine elements, among others.

4.2 Sparse storage

The discretisation matrix is a sparse matrix and storing all zeros would be very inefficient. Hence, special sparse storage formats need to be adopted. For finite difference methods, the stencils can be used to store the sparse matrix in a stencil format. In the case of a 5-point stencil, all nonzero matrix

elements come from the five elements of the stencil of each node. Hence, instead of storing the elements of the discretisation method, one could store the stencils of each node.
Now, instead of storing $N_x \times N_y$ times a stencil corresponding to a node, it is more convenient to store five times an array of $N_x \times N_y$ elements. That is, we define the arrays SN, SE, SS, SW, and SC as $(SN)_{ij} = N_{ij}$, $(SE)_{ij} = E_{ij}$, $(SS)_{ij} = S_{ij}$, $(SW)_{ij} = W_{ij}$, and $(SC)_{ij} = C_{ij}$, respectively.
There are different advantages of the stencil storage format. First and foremost, no numbering of the nodes is necessary. Second, parallelisation is easier since all necessary information for node (i, j) are stored in elements (i, j) of the stencils. Third, it can easily be extended to higher dimensions.

Remark 8. In many cases, the storage could be reduced even further. For example, stencils with constant elements could be stored as a single number. Also, for symmetric problems, the north and south stencils are the same, as well as the east and west stencils.

Exercise 19. Write down the finite difference approximation (20) in terms of coefficients of the stencil matrices SX.

Exercise 20. Show that the stencil storage is more efficient than a compressed row storage (CRS) of the discretisation matrix from the finite different method.

5 The Nyquist theory on sampling

Sampling theory deals with the question of how to represent continuous signals as discrete signals. This is related to the question on how to represent waves with approximations in discrete points. The Nyquist theory tells that you need to have at least two points in each wave period to represent a wave. Therefore, for a wavefield with wavelength λ, the spacing between points should be less than $\lambda/2$. This is a lower limit and to obtain accurate results, in practice around 10 to 20 points per wavelength are necessary in finite difference methods. For this reason, one normally fixes the number of points per wavelength and then calculates the total number of nodes that are necessary for the simulation, rather than choosing the total number of nodes directly.

Example 2. For a Helmholtz equation with a wave number of $k = 10$, we have a wavelength of $\lambda = 2\pi/k = 0.2\pi$. If we want a finite difference method with 10 points per wavelength, we need a grid spacing of $h = \lambda/10 = 0.02\pi$.

If the domain is given by the interval $[0, \pi]$, this means that $\pi/h = 50$ nodes need to be used.

Exercise 21. Derive the number of nodes necessary for a Helmholtz problem on the interval $[0, \pi]$, $k = 20$ and 10 points per wavelength.

Exercise 22. We have a Helmholtz problem on the interval $[0, \pi]$ and keep 10 points per wavelength. For a problem with wave number k_1, we need n_1 nodes. Show that for a problem of wave number $k_2 = \alpha k_1$, where α is a positive constant, the number of nodes is $n_2 = \alpha n_1$.

Exercise 23. Let us consider a two-dimensional Helmholtz problem on the rectangle $[0, 1] \times [0, 1]$. For a problem with wave number k_1, we need n_1 nodes. Now, let us consider a problem with wave number $k_2 = \alpha k_1$, where α is a positive constant. How many nodes do we need in this case? Give your answer as a function of α and n_1.

Exercise 24. In the case of high-intensity focused ultrasound treatment of liver cancer, the frequency is usually kept at 1 MHz. Propagation of sound in water has a velocity of 1500 m/s. This means that the wave length is given by $\lambda = c/f = 1500/10^6 = 1.5 \cdot 10^{-3}$, so 1.5 mm. One rib is around 15 cm long and 1.2 cm thick. We use an equidistant grid with 10 nodes per wavelength. How many nodes are necessary for one rib?

References

[1] Strang, G. *Computational Science and Engineering*. Wellesley-Cambridge Press, 2007.

[2] Burden, R.L., J.D. Faires, and A.M Burden. *Numerical Analysis*. Cengage Learning, 2015.