

# 1 Numerically solving a differential equation through a linear system

## 1.1 Abstract

In this project we aim to solve a special kind of differential equation using a numerical procedure that allows us to express the equation through a linear system. We will study some algorithms to solve such a problem, focusing on the efficiency of the program, setting our goal more on speed than generality.

The differential equation we're interested in studying is of the type

$$u''(x) = -f(x) \quad (1)$$

In our case we will limit our solutions using the contour conditions of  $u(0) = 0$  and  $u(L) = 0$ , where  $[0, L]$  is our domain of integration. Using Taylor expansion it is possible to express the second derivative of a function  $u(x)$  as

$$u''(x) = \frac{u(x-h) - 2u(x) + u(x+h)}{h^2} + \mathcal{O}(h^2) \quad (2)$$

We are therefore able to discretize equation (1) using  $N$  points, obtaining:

$$u''_i = \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = -f_i \quad i \in \{1 \dots N\}$$

Using the matrix representation, we can write equation (1) as

$$\begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & \cdots & \cdots & -1 & 2 \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix} = h^2 \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_{N-2} \\ f_{N-1} \end{pmatrix} \quad (3)$$

Note how, with this system it is already implied that  $f(0) = 0$  e  $f(L) = 0$ , since the first and last equations state that

$$h^2 f_0 = \frac{2u_0 - u_1}{h^2} = \frac{-1u_{-1} + 2u_0 - u_1}{h^2}$$

$$h^2 f_{N-1} = \frac{-u_{N-2} + 2u_{N-1}}{h^2} = \frac{-u_{N-2} + 2u_{N-1} - u_N}{h^2}$$

Since the boundary conditions of the differential equations state that  $u_{-1} = u(0) = 0$  and  $u_N = u(L) = 0$ .

This linear system is indeed very particular and has a clear pattern. We will first focus on finding a solving algorithm for a general tridiagonal matrix and after we will try to implement another program to solve this particular system with the intent of lowering the number of calculation and therefore the computation time.

## 1.2 General algorithm for solving a tridiagonal matrix through back and forward substitution

A general tridiagonal system can be expressed as

$$\begin{pmatrix} b_0 & c_0 & 0 & 0 & \cdots & 0 \\ a_1 & b_1 & c_1 & 0 & \cdots & 0 \\ 0 & a_2 & b_2 & c_2 & \cdots & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & a_{N-2} & b_{N-2} & c_{N-2} \\ 0 & 0 & \cdots & \cdots & a_{N-1} & b_{N-1} \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix} = h^2 \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_{N-2} \\ f_{N-1} \end{pmatrix} \quad (4)$$

We will describe the algorithm we used for this system first for a 3x3 tridiagonal matrix, and after we will demonstrate its validity for a square tridiagonal matrix of optional dimension.

MATRIX 3X3:

$$\left( \begin{array}{ccc|c} b & c & 0 & f_0 \\ a & b & c & f_1 \\ 0 & a & b & f_2 \end{array} \right) \longrightarrow \quad (5)$$

$$\text{Passage1} : \left( \begin{array}{ccc|c} 1 & c/b & 0 & f_0/b \\ a & b & c & f_1 \\ 0 & a & b & f_2 \end{array} \right) \longrightarrow \quad (6)$$

$$\text{Passage2} : \left( \begin{array}{ccc|c} 1 & c/b & 0 & f_0/b \\ 0 & \frac{b-(c/b)a}{b-(c/b)a} & \frac{c}{b-(c/b)a} & \frac{f_1-af_0}{b-(c/b)a} \\ 0 & 0 & \frac{b-\frac{ac}{b-(c/b)a}}{b-\frac{ac}{b-(c/b)a}} & \frac{f_2-af_1}{b-\frac{ac}{b-(c/b)a}} \end{array} \right) = \left( \begin{array}{ccc|c} 1 & c/b & 0 & f_0/b \\ 0 & 1 & \frac{c}{b-ac/b} & \frac{f_1-af_0}{b-(c/b)a} \\ 0 & 0 & 1 & \frac{f_2-af_1}{b-\frac{ac}{b-(c/b)a}} \end{array} \right) \longrightarrow \quad (7)$$

$$\text{Passage3} : \left( \begin{array}{ccc|c} 1 & 0 & 0 & f_0/b - f_1(c/b) \\ 0 & 1 & 0 & \frac{f_1-af_0}{b-(c/b)a} - f_2 \frac{c}{b-ac/b} \\ 0 & 0 & 1 & \frac{f_2-af_1}{b-\frac{ac}{b-(c/b)a}} \end{array} \right) \quad (8)$$

Now it's very simple to solve the system.

MATRIX (n+1)x(n+1)

Before starting to demonstrate that the above passages can be done also for a (n+1)x(n+1) matrix, supposed that they work for a nxn one, we can notice that, in general, for a square matrix of optional dimension N, doing the passage 2 until the penultimate row we obtain:

$$A_{N-1,N} = \frac{c}{bet(N-1)}$$

where

$$bet(n) = b_0 - \frac{ac}{b_1 - \frac{ac}{b_2 - \frac{ac}{\ddots - \frac{ac}{b_{n-1} - \frac{ac}{b_n}}}}}$$

(here all the  $b_i$ 's have the same value; the index i helps only to count them).

Now we do the passage 2 until the last row (we focus only on the tridiagonal matrix; if we manage to obtain the unitary matrix the system is solved); we obtain:

$$\left( \begin{array}{ccccc} 1 & c/b & 0 & 0 & \dots \\ 0 & 1 & \frac{c}{b-ac/b} & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & 1 & c/bet(n) \\ \dots & \dots & \dots & 0 & 1 \end{array} \right) \quad (9)$$

and simply subtracting, from the n-row, the (n+1)-row multiplied for bet(n)/c:

$$\left( \begin{array}{ccccc} 1 & c/b & 0 & 0 & \dots \\ 0 & 1 & \frac{c}{b-ac/b} & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & 1 & 0 \\ \dots & \dots & \dots & 0 & 1 \end{array} \right) \quad (10)$$

Now, ignoring the (n+1)-row and the (n+1)-column, we have a nxn matrix which we can bring back to the identity going on with the passage 3.

ALGORITHM IN C++

Translating the above algorithm in C++ language and working only on the vector f and on the solution vector u, we obtain the following code:

```

u[0]=f[0]/(bet=b);
for(int j = 1; j < N; j++) {
    gam[j]=c/bet;
    bet=b-a*gam[j];
    u[j]=(f[j]-a*u[j-1])/bet;
}
for (int j = (N-2); j >= 0; j--) u[j] -= gam[j+1]*u[j+1];

```

1  
2  
3  
4  
5  
6  
7

### 1.3 Particular algorithm

Using the regular Gaussian elimination algorithm we proceed to find a specific solution of our system as follows; we will start form a 3x3 matrix as before, and then demonstrate it for a matrix of optional dimension.

MATRIX 3X3:

$$\left( \begin{array}{ccc|c} 2 & -1 & 0 & f_0 \\ -1 & 2 & -1 & f_1 \\ 0 & -1 & 2 & f_2 \end{array} \right) \rightarrow \quad (11)$$

$$Passage1 : \left( \begin{array}{ccc|c} 2 & -1 & 0 & f_0 \\ 0 & 3 & -2 & f_1(1+1) + f_0 \\ 0 & 0 & 4 & f_2(2+1) + f_1 \end{array} \right) \quad (12)$$

that is:  $a_{i,j} \rightarrow (i+1)a_{i,j} + a_{i-1,j}$  for  $i$  going from 1 to  $(n-1)$ , where  $n$  is the matrix's dimension (3 in this case)

$$\rightarrow Passage2 : \left( \begin{array}{ccc|c} 2 & -1 & 0 & f_0 \\ 0 & 3 & -1 & f_1(1+1) + f_0 \\ 0 & 0 & 1 & \frac{f_2}{3+1} \end{array} \right) \quad (13)$$

that is:  $a_{n-1,j} \rightarrow \frac{an-1,j}{n+1}$  where  $n$  is the matrix's dimension

$$Passage3 : \rightarrow \left( \begin{array}{ccc|c} 1 & 0 & 0 & \frac{f_0+(0+1)f_1}{0+2} \\ 0 & 1 & 0 & \frac{f_1+(1+1)f_2}{1+2} \\ 0 & 0 & 1 & \frac{f_2}{3+1} \end{array} \right) \quad (14)$$

that is  $a_{i,j} \rightarrow \frac{a_{i,j}+(i+1)a_{i+1,j}}{i+2}$  for  $i$  going from 0 to  $(n-1)$

Now it's very simple to solve the system.

MATRIX  $(n+1) \times (n+1)$

We focus only on the tridiagonal matrix and try to obtain the unitary matrix; to do this let's to passage 1 until the  $n+1$  row.

$$\left( \begin{array}{ccccc} 2 & -1 & 0 & 0 & \dots \\ 0 & 3 & -2 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & n+2 & -(n+1) \\ \dots & \dots & \dots & -1 & 2 \end{array} \right) \quad (15)$$

and simply subtracting, from the  $n$ -row, the  $(n+1)$ -row multiplied for  $\text{bet}(n)/c$ :

$$\left( \begin{array}{ccccc} 1 & c/b & 0 & 0 & \dots \\ 0 & 1 & \frac{c}{b-ac/b} & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & 1 & 0 \\ \dots & \dots & \dots & 0 & 1 \end{array} \right) \quad (16)$$

Now, ignoring the  $(n+1)$ -row and the  $(n+1)$ -column, we have a  $n \times n$  matrix which we can bring back to the identity going on with the passage 3.

#### ALGORITHM IN C++

Translating the above algorithm in C++ language and working only on the vector  $f$  and on the solution vector  $u$ , we obtain the following code:

```

    u[0]=f[0]/(bet=b);
    for(int j = 1; j < N; j++) {
        gam[j]=c/bet;
        bet=b-a*gam[j];
        u[j]=(f[j]-a*u[j-1])/bet;
    }
    for (int j = (N-2); j >= 0; j--) u[j] -= gam[j+1]*u[j+1];

```

1  
2  
3  
4  
5  
6  
7

## 1.4 Errors

## 1.5 Time

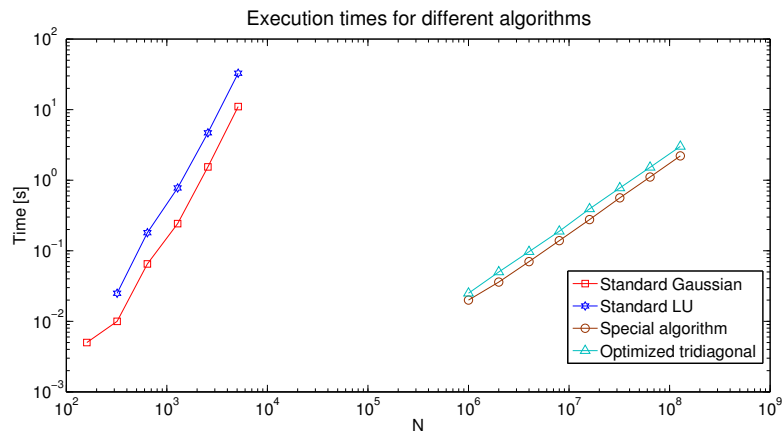


Figure 1: \*\*\*