Exercise 2

Introduction

The goal of this exercise is to discretize the differential equation with variable coefficients

$$-(a(x)u'(x))' = f(x), \qquad x \in [0,1]$$

with $0 < a(x) \leq \bar{a}$, subject to homogeneous Dirichlet boundary conditions and Neumann boundary conditions.
The first step is to discretize the domain using a mesh with step size h to identify $n+2$ nodes, and then introduce nodes $x_{j\pm1/2} = x_j \pm \frac{1}{2}h$ for $j = 1, ..., n$. The second step is to use these nodes to define a discretization scheme and write down the resulting linear system. The properties of the coefficient matrix are then investigated, and the discrete maximum principle is derived if possible.

For the case of Neumann boundary conditions, the compatibility condition is written down, and its implications for the solvability of the problem are discussed. The discretization scheme and one-sided derivatives are then used to find the discrete equivalent of the compatibility condition, and its second-order approximation is verified for the case where $a(x)$ is constant.
Finally, the accuracy of the approximations is tested using MATLAB, and the results are discussed.

Analysis

Discretization with Dirichlet boundary conditions The discretization of the differential equation is given by:

$$\frac{1}{h^2}[a_{j-1/2}u_{j-1} + (a_{j-1/2} + a_{j+1/2})u_j - a_{j+1/2}u_{j+1}] = f_j, \quad j = 1, ..., n$$

where h is the step size and $a_{j\pm1/2} = a(x_{j\pm1/2})$. According to homogeneous Dirichlet boundary conditions, we also want $u_0 = u_{n+1} = 0$. This discretization is consistent with the second-order central difference approximation for the second derivative. The corresponding linear system is $\mathbf{Au} = \mathbf{f}$, where $\mathbf{A}$ is the matrix with elements:

$$\begin{bmatrix} a_{1-1/2} + a_{1+1/2} & -a_{1+1/2} & 0 & \cdots & 0 \\ -a_{2-1/2} & a_{2-1/2} + a_{2+1/2} & -a_{2+1/2} & 0 & \cdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \cdots & -a_{(n-1)-1/2} & a_{(n-1)-1/2} + a_{(n-1)+1/2} & -a_{(n-1)+1)} \\ 0 & \cdots & 0 & -a_{n-1/2} & a_{n-1/2} + a_{n+1/2} \end{bmatrix}$$

and $\mathbf{f}$ is a vector with elements f_j.
The matrix $\mathbf{A}$ is symmetric and strict diagonally dominant. In addition, we have $a_{jj} > 0$ for each $j = 1, ..., n$, since $\min_{x\in[0,1]} a(x) > 0$; therefore $\mathbf{A}$ is positive definite and invertible which ensures the existence of a unique solution to the linear system of equations. Furthermore, the matrix $\mathbf{A}$ is tridiagonal which makes it easy to solve the system efficiently using the \ MATLAB operator. Hence, we obtain:

$$\mathbf{A} \cdot \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix}$$

Dirichlet code summary Let's get into detail with an example, using $a(x) = e^{-x}$ and $f(x) = 1$, and computing results with MATLAB.
We first use the MATLAB symbolic toolbox and the `dsolve` command to find the exact solution to our Dirichlet problem.

```
syms u(x) a(x) f(x)
Du=diff(u,1);
f=1;
a=exp(-x);
uex=dsolve(-diff((a*Du),1)==f,u(0)==0,u(1)==0);
uex=simplify(uex)
```

We immediately find our exact solution to be $u(x) = e^x(\frac{1}{e-1} - x + 1) - \frac{e}{e-1}$.
We now want to use our discretization model discussed above to find the approximate solution to our problem and compare it with the real one.
Let's follow the MATLAB script `proj2dir.m`, starting with 12 nodes (10 internal nodes + 2 boundaries nodes, `n=10`) and mesh step `h=1/11`. In rows `12-22`, we have some initial definitions and initialization. We also specify Dirichlet boundary conditions. We proceed to create the coefficient matrix using the `spdiags()` function. We define the diagonal and off-diagonal values of the matrix according to the discretization scheme given above. We modify the known term vector `b` to account for the boundary conditions by adding the values at the first and last nodes, which correspond to the boundary values, to the appropriate entries of the vector. Finally, we solve the linear system using the backslash operator and obtain the numerical solution `Uh`, which we plot along with the exact solution `uex`.

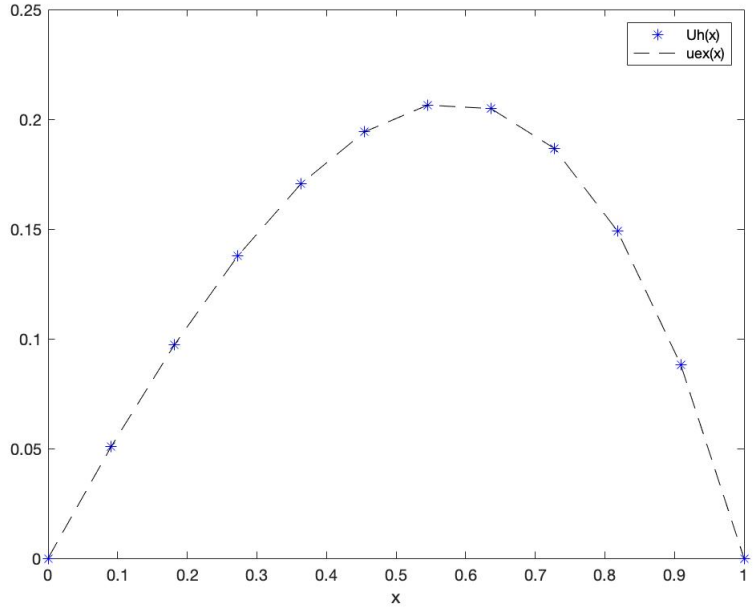

Figure 4: Comparison between exact and approximated solution, with 12 nodes.

As we can see, even considering a small number of nodes, the value of the approximation on each node lies close to the curve of the exact solution. To verify the accuracy of the numerical solution, we can calculate the error norm between the numerical and exact solutions and output the value of h and the error norm.
In the example above, we find `h = 9.0909090909e-02 err = 7.1067865137e-05`, which can be considered accurate enough. We repeat the same procedure for different numbers of nodes to verify the order of accuracy of the numerical scheme. The results are shown in the following table:

Number of spacings (`n`)	Mesh step (`h`)	Error norm (`err`)
10	9.0909e-02	7.1068e-05
20	4.7619e-02	1.9603e-05
50	1.9608e-02	3.3250e-06
100	9.9010e-03	8.4792e-07
200	4.9751e-03	2.1410e-07
500	1.9960e-03	3.4461e-08

Finally, we can verify the maximum principle using the following results:

Theorem. *(Discrete Maximum Principle)*
Let's consider the following ODE: $-\frac{d}{dx}(p(x)\frac{du}{dx}+q(x)u=f(x)$ *for* $x\in(a,b)$.
Let L_h *be our discretization method.*
(i) If $L_hu_j\leq 0$ *for all* $1\leq j\leq n$ *and* $\max(u_0,u_{n+1})\geq 0$, *then* $\max_{0\leq j\leq n+1}u_j\leq\max(u_0,u_{n+1})$.
(ii) If $L_hu_j\geq 0$ *for all* $1\leq j\leq n$, *and* $\min(u_0,u_{n+1})\leq 0$, *then* $\min_{0\leq j\leq n+1}u_j\geq\min(u_0,u_{n+1})$.

Corollary. *If* $q(x)\equiv 0$, *then the discrete maximum principle holds without the hypotheses* $\max(u_0,u_{n+1})\geq 0$ *for part (i) and* $\min(u_0,u_{n+1})\leq 0$ *for part (ii).*

Given that the value of L_hu_j is exactly $f_j=f(x_j)$ for each $j=1,...,n$, where the function $f(x)=e^{-x}$, therefore always positive, the hypothesis are satisfied.

Discretization with Neumann boundary conditions We now want to solve our differential problem, considering the following Neumann boundary conditions:

$$u'(0)=g_0 \qquad u'(1)=g_1$$

with g_0 and g_1 assigned real numbers. We use the same discretization scheme seen above:

$$\frac{1}{h^2}[a_{j-1/2}u_{j-1}+(a_{j-1/2}+a_{j+1/2})u_j-a_{j+1/2}u_{j+1}]=f_j,\quad j=1,...,n$$

However, u_0 and u_{n+1} can't be determined from the boundary conditions, therefore they will be added to the unknown values, components of the vector $\mathbf{u}$.
To incorporate the boundary conditions into the discretization scheme, we use the one-sided finite difference approximations for $u'(0)$ and $u'(1)$, given as:

$$u'(0)\simeq\frac{1}{h}\left(-\frac{3}{2}u(0)+2u(h)-\frac{1}{2}u(2h)\right)$$

$$u'(1)\simeq\frac{1}{h}\left(\frac{1}{2}u(1-2h)-2u(1-h)+\frac{3}{2}u(1)\right)$$

The corresponding linear system is $\mathbf{Au}=\mathbf{f}$, where $\mathbf{A}$ is the matrix with elements:

$$\begin{bmatrix} -3/2 & 2 & -1/2 & 0 & \cdots & 0 \\ -a_{1-1/2} & a_{1-1/2}+a_{1+1/2} & -a_{1+1/2} & 0 & \cdots & 0 \\ 0 & -a_{2-1/2} & a_{2-1/2}+a_{2+1/2} & -a_{2+1/2} & 0 & \cdots \\ 0 & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \cdots & 0 & -a_{n-1/2} & a_{n-1/2}+a_{n+1/2} & -a_{n+1/2} \\ 0 & \cdots & \cdots & 1/2 & -2 & 3/2 \end{bmatrix}$$

and $\mathbf{f}$ is a vector with elements f_j.
Hence, we obtain:

$$\mathbf{A} \cdot \begin{bmatrix} u_0 \\ \vdots \\ \vdots \\ u_{n+1} \end{bmatrix} = \begin{bmatrix} g_0/h \\ f_1 \\ \vdots \\ f_n \\ g_1/h \end{bmatrix}$$

The values $g0/h$ and $g1/h$ in the known terms vector result from the incorporation of discretized Neumann boundary conditions into our system of equations. To verify the accuracy order of the one-sided derivative formula for $u'(0)$, we need to perform a Taylor series expansion of $u(x)$ around $x = 0$. We have:

$$u(h) = u(0) + hu'(0) + \frac{h^2}{2}u''(0) + o(h^3)$$

$$u(2h) = u(0) + 2hu'(0) + 2h^2u''(0) + o(h^3)$$

Substituting these expressions into the one-sided derivative formula, we get:

$$\begin{aligned} u'(0) &\simeq \frac{1}{h}\left(-\frac{3}{2}u(0) + 2u(h) - \frac{1}{2}u(2h)\right) \\ &= \frac{1}{h}\left(-\frac{3}{2}u(0) + 2u(0) + 2hu'(0) + h^2u''(0) - \frac{1}{2}u(0) - hu'(0) - h^2u''(0) + o(h^3)\right) \\ &= u'(0) + o(h^2) \end{aligned}$$

Therefore, the error of the formula for $u'(0)$ is given by:

$$u'(0) - \frac{1}{h}\left(-\frac{3}{2}u(0) + 2u(h) - \frac{1}{2}u(2h)\right) = o(h^2)$$

which implies that the one-sided derivative formula for $u'(0)$ has an accuracy of $o(h^2)$.
Similarly, for $u'(1)$ we have:

$$u(1-h) = u(1) - hu'(1) + \frac{h^2}{2}u''(1) + o(h^3)$$

$$u(1-2h) = u(1) - 2hu'(1) + 2h^2u''(1) + o(h^3)$$

Substituting these expressions into the one-sided derivative formula, we get:

$$\begin{aligned} u'(1) &\simeq \frac{1}{h}\left(\frac{3}{2}u(1) - 2u(1-h) + \frac{1}{2}u(1-2h)\right) \\ &= \frac{1}{h}\left(\frac{3}{2}u(1) - 2u(1) + 2hu'(1) - h^2u''(1) + \frac{1}{2}u(1) - hu'(1) + h^2u''(1) + o(h^3)\right) \\ &= u'(1) + o(h^2) \end{aligned}$$

Therefore, the error of the formula for $u'(1)$ is given by:

$$u'(1) - \frac{1}{h}\left(\frac{3}{2}u(1) - 2u(1-h) + \frac{1}{2}u(1-2h)\right) = o(h^2)$$

which implies that the one-sided derivative formula for $u'(1)$ has again an accuracy of $o(h^2)$.
To guarantee the existence of a unique solution, we have to satisfy the *compatibility condition*.

The *compatibility condition* for the differential equation $-(a(x)u'(x))' = f(x)$ can be obtained by integrating the differential equation over the domain $[0,1]$ and using integration by parts:

$$\begin{aligned}
\int_0^1 -\left(a(x)u'(x)\right)' dx &= \int_0^1 f(x)dx \\
-\left[a(x)u'(x)\right]_0^1 &= \int_0^1 f(x)dx \\
a(0)u'(0) - a(1)u'(1) &= \int_0^1 f(x)dx \\
g_0 a(0) - g_1 a(1) &= \int_0^1 f(x)dx \qquad (*)
\end{aligned}$$

where we have used the boundary conditions $u'(0) = g_0$ and $u'(1) = g_1$ in the third line.
Considering $n+2$ nodes and $a(x) = \bar{a} \in \mathbb{R}$, we obtain the following discrete version of $(*)$:

$$\bar{a}(g_0 - g_1) = \sum_{j=0}^{n} f(x_j) \frac{1}{n+1} = h \sum_{j=0}^{n} f(x_j) \quad (**)$$

where g_0 and g_1 are the one-sided derivatives discussed above. Since g_0 and g_1 are second-order approximations and $\bar{a}$ is a constant, the $(**)$ is also a second-order approximation of $(*)$.

Neumann code summary Let's get into detail using again $a(x) = e^{-x}$ and $f(x) = 1$, and computing results with MATLAB.
As we did for Dirichlet BCs, we use the MATLAB symbolic toolbox and the `dsolve` command to find the exact solution to our Neumann problem. The algorithm can be found on the `exactsol.m` MATLAB script.

```
syms u(x) a(x) f(x)
Du=diff(u,1);
f=1;
a=exp(-x);
temp_uex=dsolve(-diff((a*Du),1)==f);
temp_uex=simplify(temp_uex);
% uex=vectorize(uex)
temp_uex
temp_duex=diff(temp_uex,1)
der_value_in0=subs(temp_duex,x,0)
C1=0 % value of integration constant. Can be seen as u'(0).
uex = exp(x)*(C1 - x + 1);
duex=diff(uex,1)
der_value_in1=subs(duex,x,1)
```

We arbitrarily choose the value of g_0 (`C1` in the script) to be 0. Our computation shows how the value of g_1 is then forced to be $-e$. Using our analytic results, from $(*)$ we have:

$$g_0 e^0 - g_1 e^{-1} = \int_0^1 1 dx$$

hence, $g_1 = (g_0 - 1)e = (0-1)e = -e$, as found using MATLAB.
From now on, we will assume $(g_1, g_2) = (0, -e)$.
We find our exact solution to be $u(x) = e^x(1-x)$. We now want to use the discretization discussed above to find the approximate solution to our Neumann problem and compare it with the real one.

Let's follow the MATLAB script `proj2neum.m`, starting with 12 nodes (10 internal nodes + 2 boundaries nodes, `n=10`) and mesh step `h=1/11`. In rows `19-29`, we have some initial definitions and initialization. We proceed to create the coefficient matrix using the `spdiags()` function. We define the diagonal and off-diagonal values of the matrix according to the discretization scheme given above. This time, we will have two extra rows that incorporate the equation to represent the one-sided approximation of $u'(0) = g_0 = 0$ and $u'(1) = g_1 = -e$. We modify the known term vector `b` to account for the boundary conditions by adding the values g_0/h and g_1/h to the appropriate entries of the vector. Finally, we solve the linear system using the backslash operator and obtain the numerical solution `Uh`, which we plot along with the exact solution `uex`.

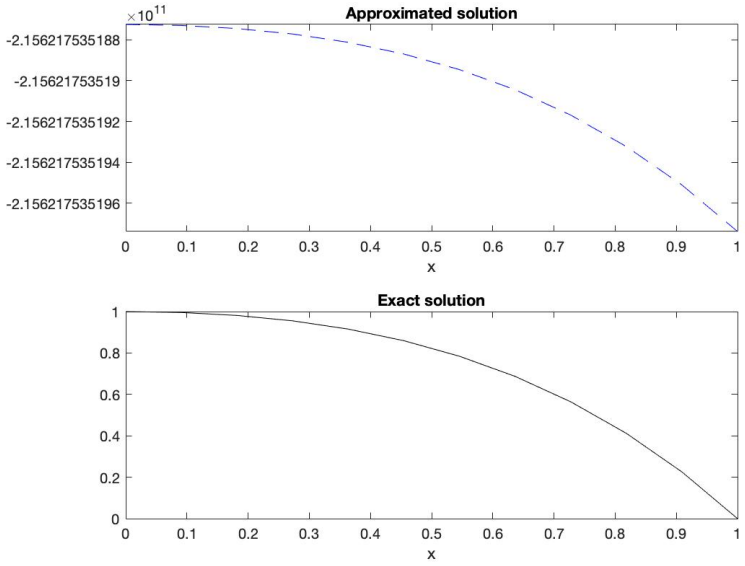

Figure 5: Comparison between exact and approximated solution, with 12 nodes.

We can easily notice how our solutions both respect the Neumann boundaries condition. However, they differ from a constant. In fact, the generic solution to our problem is $C_2 + e^x(C_1 - x + 1)$. By imposing the Neumann BCs - differentiating the solution - we lose the information on the additive constant C_1. This results in an infinite set of solutions that satisfy our boundary problem, even though they are all translated by a constant.
To have a significantly comparable pair of solutions, we need to add a Dirichlet boundary condition in $x = 0$. Results of this implementation can be found on the MATLAB script `proj2neum_dirichlet.m`. Plotting both the approximated and exact solution, we now have:

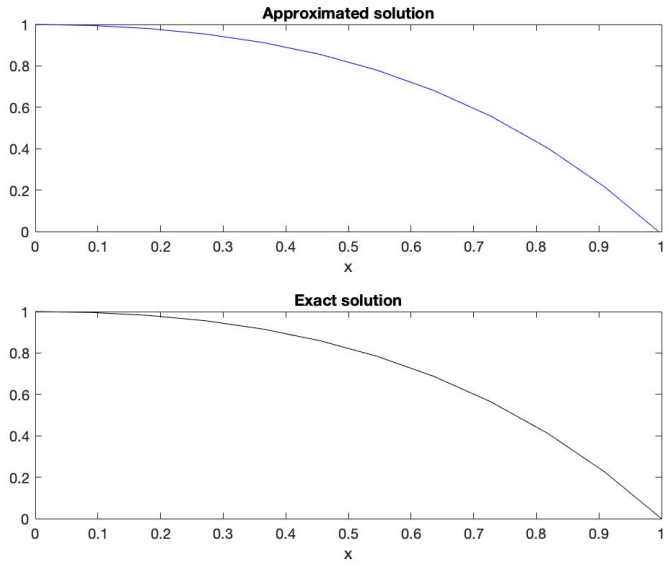

Figure 6: Comparison between exact and approximated solution, with Neumann+Dirichlet BCs.

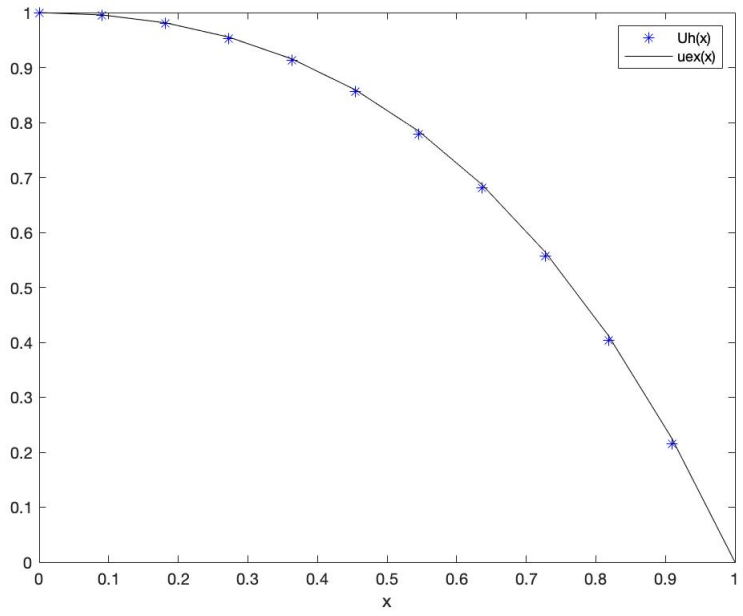

Figure 7: Comparison between exact and approximated solution, with 12 nodes highlighted.

To verify the accuracy of the numerical solution, we can calculate the error norm between the numerical and exact solutions as we did with the Dirichlet problem, and output the value of h and the error norm. In the example above, we find `h = 9.0909090909e-02 err = 1.2066273753e-02`, which can be considered accurate enough. We repeat the same procedure for different numbers of nodes to verify the order of accuracy of the numerical scheme. The results are shown in the following table:

Number of spacings (`n`)	Mesh step (`h`)	Error norm (`err`)
10	9.0909e-02	1.2066e-02
20	4.7619e-02	3.3268e-03
50	1.9608e-02	5.6642e-04
100	9.9010e-03	1.4468e-04
200	4.9751e-03	3.6566e-05
500	1.9960e-03	5.8893e-06

Conclusions

In conclusion, the proposed discretization method appears to be quite accurate. However, the accuracy of the method was found to be much higher in the Dirichlet boundary problem, compared to the Neumann boundary problem. For instance, when considering a node number of $100 + 2$ ($\mathtt{n} = 100$), in the former case, the error is of the order of 10^{-7}, whereas in the latter case, it is of the order of 10^{-4}.