ABSTRACT. Many science and engineering problems can be mathematically modeled as a boundary value problem. Thus, it is very important to build a method to solve for a boundary value problem. For some simple boundary value problems, for example, a two-dimensional Laplace equation, it may be possible to solve it analytically by separation of variables. However, in most applications, boundary value problems are much more complex and there are no available analytical methods. For this reason, a numerical method called finite element method (FEM) is developed for the analysis of boundary value problems. It approximates the domain by many small elements and simplify possible integrations. In this paper, FEM is applied to solve a boundary value problem of a one-dimensional Helmholtz equation. It can be seen that FEM provides a fast and accurate approach for this problem.

1. INTRODUCTION

In this paper, a finite element analysis is developed for a boundary value problem of a one-dimensional Helmholtz equation. We follow the presentation of the Finite Element Method (FEM for short), using piece-wise linear elements, in the book by T. J. R. Hughes [2]. The finite element method (FEM) uses the weak form of the Helmholtz equation. Based on the weak form, a finite-dimensional representation of the solution is obtained from a basis of simple functions, here “hat functions” (piece-wise linear functions) that are localized to sufficiently small elements of the domain for the problem. This approach makes it possible to solve the differential equations in a matrix form. We will compare the result from FEM is compared with the analytical solution in a simple case. The accuracy of FEM can therefore be evaluated from the comparison.

Date: October 22, 2014.
1.1. Definition of the boundary value problem and weak form.
The Helmholtz equation considered in this paper is a second order, non-
homogeneous partial differential equation defined in a one-dimensional
domain, $\Omega$. The partial differential equation can be expressed as:

$$
\frac{d^2 u}{dx^2} + ku(x) = l(x)
$$

where $u$ is a function in $\Omega$, $k$ is a given coefficient, $l$ is an arbitrary
function in $\Omega$ and $\Omega$ is a one-dimensional interval, $[0, 1]$. The boundary
values are specified as:

$$
\begin{align*}
    u(0) &= 0 \\
    u(1) &= 0
\end{align*}
$$

The purpose is to solve the function $u$ in $\Omega$. To achieve this purpose
by FEM, the original form of the partial differential equation implies
no useful information for domain discretization. Therefore, the original
equation needs to be transformed to a new format, called weak form.
To generate the weak form, assume that there is a test function $w$
which follows the same boundary value conditions as $u$. Moreover, the
basis of $w$ is the same as $u$. If multiply $w$ to equation (1) and take
integration to both sides, equation (1) can be changed to:

$$
\int_0^1 w \frac{d^2 u}{dx^2} dx + k \int_0^1 w u dx = \int_0^1 w l dx
$$

Apply integral by parts to equation (2), get:

$$
\left[ w \frac{du}{dx} \right]_0^1 - \int_0^1 \frac{dw}{dx} \frac{du}{dx} dx + k \int_0^1 w u dx = \int_0^1 w l dx
$$

which is the weak form of the boundary value problem.

1.2. Equation discretization. The weak form is an integral equation
of function $w$, $u$ and $l$. The format of the equation implies that if it is
possible to choose an available basis for function $w$ and $u$ in $\Omega$, $w$ and $u$
can be represented by the sum of the elements of the basis. Thus, the
equation is discretized in the domain and the integrals can be solved
easily for each element. To see the process, assume one element of the
basis is $N$ such that:

$$
\begin{align*}
    w \approx w^h &= \sum_{A=1}^{n} c_A N_A \\
    u \approx u^h &= \sum_{B=1}^{n} d_B N_B \\
    N_A(0) &= N_A(1) = 0 \\
    N_B(0) &= N_B(1) = 0
\end{align*}
$$

(4)

In theory, if the value of $n$ approaches infinity, $w = w^h$ and $u = u^h$. However, in real calculations, it is impossible to use an infinite value. Nevertheless, it is still expected that as long as the value of $n$ is large enough, the function $u^h$ and $u^h$ should be good approximations with small errors. With the function $w^h$ and $u^h$ defined and replaced with $w$ and $u$ in equation (3), get a sum of integral equations of the elements of basis:

$$
\sum_{B=1}^{n} d_B \left[ \int_0^1 N_A N_B dx - k \int_0^1 N_A N_B dx \right] = - \int_0^1 N_A ldx
$$

(5)

Equation (5) is just a system of linear equations and implies the use of matrix algebra.

1.3. **The basis of the function and matrix formulation.** For convenience, define

$$
\begin{align*}
    S_{AB} &\equiv \int_0^1 N'_A N'_B dx \\
    P_{AB} &\equiv k \int_0^1 N_A N_B dx \\
    F_A &\equiv \int_0^1 N_A ldx
\end{align*}
$$

Thus, equation (5) can be represented as:

$$
\sum_{B=1}^{n} (S_{AB} + P_{AB}) d_B = F_A
$$

(6)

Moreover, matrix $S$, $P$ and $K$, vector $F$, and $d$ can be defined as:

$$
S = [S_{AB}] = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\
S_{21} & S_{22} & \cdots & S_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
S_{n1} & S_{n2} & \cdots & S_{nn} \end{bmatrix}
$$

$$
P = [P_{AB}] = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\
P_{21} & P_{22} & \cdots & P_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
P_{n1} & S_{n2} & \cdots & P_{nn} \end{bmatrix}
$$

$$
K = S - P
$$
\[
F = [F_A] = \begin{bmatrix}
F_1 \\
F_2 \\
\vdots \\
F_n
\end{bmatrix}
\]
\[
d = [d_A] = \begin{bmatrix}
d_1 \\
d_2 \\
\vdots \\
d_n
\end{bmatrix}
\]

With those matrices and vectors defined, the linear equation system represented by equation (6) can be solved by matrix algebra:

(7) \[ KD = F \]

2. Basis determination and calculation of integrals

For the problem of a one-dimensional Helmholtz equation, the basis of the test function can be chosen as hat functions. To formulate appropriate hat functions, discretize the interval \([0, 1]\) to \(2^n\) elements of length, \(h = \frac{1}{2^n}\). Label the start point of the elements as \(x_1, x_2, x_3 \ldots x_{2^n}\) and label the point \(x = 1\) (which is just the end point of the last element) as \(x_{2^n+1}\). The hat functions, \(N_A\) are therefore defined as:

\[
N_A(x) = \begin{cases}
\frac{(x-x_A)}{h}, & x_A \leq x \leq x_{A+1} \\
\frac{(x_{A+2}-x)}{h}, & x_{A+1} \leq x \leq x_{A+2} \\
0, & \text{elsewhere}
\end{cases}
\]

For example, if \(n = 2\), the hat functions are shown in Figure (1). An important point should be noticed is that the value of \(A\) is smaller than the number of elements. Thus, the definition of \(n\) in this section is in fact different from the definition of \(n\) in the introduction section. In another word, the value of \(n\) in this section is equal to \(n + 1\) for the value of \(n\) in the introduction section.

With the hat functions defined, the basis is well determined. Thus, the integrals in equation (6) can be calculated and therefore, the entries in matrices \(S\) and \(P\), vector \(F\) are determined. Due to the properties of the hat functions, it turns out that most entries in the matrices are 0 and the remained entries are only related to the value of \(h\) and \(k\),
which is shown as following:

\[
S = \frac{1}{h} \begin{bmatrix}
2 & -1 & 0 & \cdots & 0 \\
-1 & 2 & -1 & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
0 & \cdots & -1 & 2 \\
\end{bmatrix}
\]

\[
P = \frac{h}{6k} \begin{bmatrix}
4 & 1 & 0 & \cdots & 0 \\
1 & 4 & 1 & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
0 & \cdots & 1 & 4 \\
\end{bmatrix}
\]

Therefore, the matrix \( K \) can be easily generated from these matrices.

The generation of vector \( F \) is a little more complex. Since the integral for the calculation of \( F \) involves an unknown function \( l \), it cannot be guaranteed that the vector can be expressed with a simple format as the \( K \) matrix. Numerical integration has to be applied to calculate the entries.
3. Case study

Because the size of the matrices involved in the FEM needs to be large to achieve reasonable results, it is difficult to solve the matrix manually even though the calculation method is simple. Moreover, there are integrations needs to be solved numerically. Due to the two points, it is preferable to use computer for efficient calculation. MATLAB is a computing language which can be used for this task.

In this section, the function \( l(x) \) is defined as \( l(x) = -\sin(\pi x) \). The analytical and numerical results for different values of \( k \) is compared and discussed.

3.1. Analytical solution. To solve for equation (1), assume that the function \( u \) can be expressed as Fourier series, such that:

\[
u(x) = \sum_{m=0}^{\infty} [a_m \sin(m\pi x) + b_m \cos(m\pi x)]
\]

Combine equation (1) and (8), get

\[
-\sum_{m=0}^{\infty} n^2 \pi^2 [a_m \sin(m\pi x) + b_m \cos(m\pi x)] + k\sum_{m=0}^{\infty} a_m \sin(m\pi x) = -\sin(\pi x)
\]

From equation (9), it can be seen that \( m \) only equals 1. Thus, the analytical solution is

\[
u(x) = \frac{1}{\pi^2 - k} \sin(\pi x)
\]

3.2. FEM solution and results comparison. Results are calculated by FEM for different values of \( k \). For a direct comparison, the analytical solution and numerical solution are presented in figure 2 for \( k = 0.5\pi^2 \) and \( n = 2, n = 4 \) respectively. The symbol \( n \) is defined as the same way in section 2.

For a better comparison, errors of the numerical solution are tabulated in table 1. The errors are defined as the absolute value of the maximum difference between the numerical and analytical solution.

From table 1, it can be seen that for same value of \( k \), the error becomes smaller for larger \( n \). That is reasonable since larger \( n \) means more elements and therefore implies better accuracy. However, that does not mean that larger \( n \) necessarily achieves better result. From the table, it is clear that when the value of \( n \) is very large, the result can be worse than a smaller value of \( n \).

Another point is that for the value of \( k \) closer to \( \pi^2 \), the error tends to be larger. From equation (10), it is clear that when \( k = \pi^2 \), there will be not a good defined analytical solution. Thus, it is expected that when the value of \( k \) is close to \( \pi^2 \), the error should be larger. If
Figure 2. Analytical solution and numerical solution for $k = 0.5\pi^2$

Table 1. FEM error

<table>
<thead>
<tr>
<th>n</th>
<th>$k = 0$</th>
<th>$k = 0.5\pi^2$</th>
<th>$k = 0.9\pi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.00713</td>
<td>0.00714</td>
<td>0.347</td>
</tr>
<tr>
<td>3</td>
<td>0.00191</td>
<td>0.00191</td>
<td>0.120</td>
</tr>
<tr>
<td>4</td>
<td>0.000491</td>
<td>0.000491</td>
<td>0.0331</td>
</tr>
<tr>
<td>5</td>
<td>0.000127</td>
<td>0.000127</td>
<td>0.00852</td>
</tr>
<tr>
<td>6</td>
<td>3.56e-05</td>
<td>3.56e-05</td>
<td>0.00218</td>
</tr>
<tr>
<td>7</td>
<td>1.27e-05</td>
<td>1.27e-05</td>
<td>0.000584</td>
</tr>
<tr>
<td>8</td>
<td>6.97e-06</td>
<td>6.97e-06</td>
<td>0.000184</td>
</tr>
<tr>
<td>9</td>
<td>5.54e-06</td>
<td>5.54e-06</td>
<td>8.40e-05</td>
</tr>
<tr>
<td>10</td>
<td>5.19e-06</td>
<td>5.19e-06</td>
<td>8.40e-05</td>
</tr>
</tbody>
</table>

the value of $k$ is exactly $\pi^2$, there should not be a solution for FEM. To verify this point, the FEM is tested under the condition of $k = \pi^2$. The computer is still able to generate some results and the results are presented in figure 3.

From figure 3, the results never converge. It implies that FEM fails when $k = \pi^2$ even though the computer can still calculate some wrong results.

In fact, it is not necessary to make $k = \pi^2$ to generate a wrong result; however, if the value of $k$ is too close to $\pi^2$, the result can still be not correct. To present this point, figure 4 are the results generated for
$k = 1.00001\pi^2$. Obviously, unlike figure 2, the results in figure 4 do not converge smoothly to the exact solution.

Furthermore, the time used for all calculations for this case study is within 1 minute. The computer used for the calculation is only a typical commercial PC, a fact which implies that FEM is an efficiency numerical method for a boundary value problem of a one-dimensional Helmholtz equation.

The accuracy of the FEM is limited by the accuracy of the numerical integration. Moreover, the efficiency of the FEM is also limited by
the speed of numerical calculation. Thus, to increase the accuracy and the efficiency of the FEM, it is preferred to apply a better numerical integration method. In this case study, the numerical integration is accomplished by the built in function "trapz" in MATLAB. It approximates an integration by the area under the curve of a function. It is very straightforward but not a very accurate approach.

4. Conclusion

FEM provides a relatively efficient approach for boundary value problems of a one-dimensional Helmholtz equation. The errors behave as expected and are acceptable.

There is still not a possible approach to solve the problem for \( k = \pi^2 \) through the FEM demonstrated in the report.

To increase the efficiency and accuracy of FEM, a better method for numerical integration needs to be applied.

References
