Balancing truncation and round-off errors in practical FEM: one-dimensional analysis

Jie Liu[a,*], Matthias Möller[a], Henk M. Schuttelaars[a]

[a] *Delft Institute of Applied Mathematics*
Delft University of Technology
Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands

Abstract

In finite element methods, the solution accuracy cannot be improved indefinitely because of the limited computer precision. We propose an innovative method to find the highest attainable accuracy determined by the round-off error, for the one-dimensional second-order ordinal differential equations. This method uses a priori formula for the error evolution, so that it saves several computations on finer grids. The application of our method to a complex-valued Helmholtz equation in space shows that the highest attainable accuracy can be accurately predicted, while the required CPU time is remarkably saved.

Keywords: Finite Element Method (FEM), ordinary differential equation, round-off error, highest attainable accuracy, estimation.

1. Introduction

Many problems in engineering sciences and industry are modelled mathematically by initial-boundary value problems comprising systems of coupled, nonlinear partial or ordinary differential equations. These problems often consider complex geometries, with initial or boundary conditions that depend on measured data [1]. In some applications, not only the solution, but also its derivatives are of interest [1, 2]. For many problems of practical interest, analytical or semi-analytical solutions are not available, and hence one has to resort to numerical solution methods, such as finite difference, volume and element methods, in which the last will be adopted throughout this paper.

The accuracy of the numerically obtained solution is influenced by many sources of errors [3]: firstly, errors in the set-up of models, such as the simplification of the realistic domain and governing equations and approximation of initial and boundary conditions; next, truncation errors due to the discretization of the computational domain and use of basis functions for the function spaces defined on it; then, round-off

*Corresponding author
Email addresses: j.liu-5@tudelft.nl (Jie Liu), m.moller@tudelft.nl (Matthias Möller), h.m.schuttelaars@tudelft.nl (Henk M. Schuttelaars)

errors due to the adoption of finite-precision computer arithmetics, rather than exact arithmetics; finally, iteration errors resulting from the artificially controlled tolerance of iterative solvers.

We focus on the error led by the truncation and round-off. One tacitly assumes that the two types of errors are well-balanced. That is, the latter is often ignored based on the argument that it will be 'sufficiently small' if just IEEE-754 double-precision floating-point arithmetics are adopted. However, with the popularity of high-order approximations, the round-off error is likely to play a role with only a small number of degrees of freedom (DoFs) [4, 5, 6]. Despite this alarming observation, to the authors' best knowledge, only very few publications address the impact of accumulated round-off errors on the overall accuracy of the final solution or take them into account explicitly in the error-estimation procedure. The general rule of thumb is still to perform as many h-refinements as possible considering the available computer hardware.

The aim of this paper is to systematically analyze the influence of round-off on the error, and to propose a practical approach for obtaining the highest attainable accuracy determined by the round-off error ($E_{\min}$). The scope is restricted to one-dimensional second-order model problems using both the standard finite element method (FEM) and mixed FEM [7]. We consider the solution and its first and second derivatives, assuming the second derivative exists in the weak sense [8].

The paper is organized as follows. The model problem, finite element formulation and numerical implementation are described in Section 2. The approach to predicting $E_{\min}$ is discussed in Section 3. The parameters used in the approach are determined in Section 4. An algorithm for realizing the approach and its application to one example are put forward in Section 6. The conclusions are drawn in Section 7.

2. Model problem, finite element formulation and numerical implementation

2.1. Model problem

Consider the following one-dimensional second-order differential equation:

$$-\left(d(x)u_x\right)_x + r(x)u(x) = f(x), \qquad x \in I = [0,1], \tag{1}$$

with u denoting the unknown variable, which can either be real or complex, $f(x) \in L^2(I)$ a prescribed right-hand side, and $d(x)$ and $r(x)$ continuous coefficient functions. By choosing $d(x) = 1$ and $r(x) = 0$, Eq. (1) reduces to the Poisson equation; for $d(x) > 0$ and not constant, the diffusion equation is found when $r(x) = 0$, and the Helmholtz equation [9] is found when $r(x) \neq 0$. The boundary conditions are $u(x) = g(x)$ on Γ_D and $d(x)u_x = h(x)$ on Γ_N, where Γ_D and Γ_N are the boundaries where Dirichlet and Neumann boundary conditions are imposed, respectively.

2.2. Finite element formulation

For convenience, we introduce two inner products [10]:

$$\langle f_1,\, f_2\rangle = \int_I f_1(x) f_2(x)\, dx, \tag{2a}$$

$$\langle f_1,\, f_2\rangle_\Gamma = f_1(x_0) f_2(x_0). \tag{2b}$$

where $f_1(x)$ and $f_2(x)$ are continuous functions defined on I, and x_0 the coordinate of the boundary Γ.

2.2.1. The standard FEM

The weak form of Eq. (1) is derived in Appendix A.1. Imposing the Dirichlet boundary conditions strongly, the weak form reads:

Weak form 1

Find $u \in H^1_D(I)$ such that:

$$\langle \eta_x,\, du_x\rangle + \langle \eta,\, ru\rangle = \langle \eta,\, f\rangle + \langle \eta,\, hn\rangle_{\Gamma_N} \qquad \forall \eta \in H^1_{D0}(I),$$

with

$$H^1_D(I) = \{t \mid t \in H^1(I),\ t = g \text{ on } \Gamma_D\},$$

$$H^1_{D0}(I) = \{t \mid t \in H^1(I),\ t = 0 \text{ on } \Gamma_D\},$$

where n is 1 at $x = 1$, and -1 at $x = 0$. (3)

Imposing the Dirichlet boundary conditions in the weak sense [11], the weak form reads:

Weak form 2

Find $u \in H^1(I)$ such that:

$$\langle \eta_x,\, du_x\rangle + \langle \eta,\, ru\rangle - \langle \eta,\, du_x n\rangle_{\Gamma_D} + \langle \eta_x,\, un\rangle_{\Gamma_D} - \langle \eta,\, \rho un\rangle_{\Gamma_D}$$

$$= \langle \eta,\, f\rangle + \langle \eta,\, hn\rangle_{\Gamma_N} + \langle \eta_x,\, gn\rangle_{\Gamma_D} - \langle \eta,\, \rho gn\rangle_{\Gamma_D} \qquad \forall \eta \in H^1(I),$$

where ρ is a positive value that serves as the penalty parameter. (4)

Note that, the terms in the right-hand sides of Eqs. (3)–(4) consist of information of Neumann boundary conditions which vanish if they are not prescribed. We approximate u by a linear combination of a finite number of basis functions:

$$u \approx u_h^{(p)} = \sum_{i=1}^{m} u_i \varphi_i^{(p)}. \tag{5}$$

Here, m is the number of DoFs, $\varphi_i^{(p)}$ are C^0-continuous Lagrange basis functions supported by Gauss-Lobatto points, u_i are the values of $u_h^{(p)}$ at the DoFs, and p is the element degree. The resulting linear system of equations reads

$$AU = F, \tag{6}$$

where A is the stiffness matrix, F the right-hand side and U the discretized u.

2.2.2. The mixed FEM

Derived in Appendix A.2, the weak form of Eq. (1) using the mixed FEM is given by:

Weak form 3

Find $v \in H^1_N(I)$ and $u \in L^2(I)$ such that:

$$\langle w,\, d^{-1}v\rangle - \langle w_x,\, u\rangle = -\langle w,\, gn\rangle_{\Gamma_D} \qquad \forall w \in H^1_{N0}(I), \tag{7a}$$

$$-\langle q,\, v_x\rangle - \langle q,\, ru\rangle = -\langle q,\, f\rangle \qquad \forall q \in L^2(I), \tag{7b}$$

with

$$H^1_N(I) = \{t \mid t \in H^1(I),\ t = -h \text{ on } \Gamma_N\},$$

$$H^1_{N0}(I) = \{t \mid t \in H^1(I),\ t = 0 \text{ on } \Gamma_N\}.$$

We approximate v and u by:

$$v \approx v_h^{(p)} = \sum_{i=1}^{n} v_i \varphi_i^{(p)}, \tag{8a}$$

$$u \approx u_h^{(p-1)} = \sum_{j=1}^{p} u_{tj} \psi_j^{(p-1)},\ \text{in cell } t, \tag{8b}$$

where n is the number of DoFs for $v_h^{(p)}$, $\varphi_i^{(p)}$ are of the same type of basis functions used in Eq. (5), and v_i are the values of v_h at the DoFs; $\psi_j^{(p-1)}$ are discontinuous Lagrange basis functions and u_{tj} are the values of $u_h^{(p-1)}$ at the DoFs. The resulting coupled linear system of equations that has to be solved reads:

$$\begin{bmatrix} M & B \\ B^\top & 0 \end{bmatrix} \begin{bmatrix} V \\ U \end{bmatrix} = \begin{bmatrix} G \\ H \end{bmatrix}, \tag{9}$$

where the mass matrix M, discrete gradient operator B, and its transpose, the discrete divergence operator $B^\top$, comprise the left-hand side; G and H are the components of the right-hand side; V and U are the discretized v and u.

For the sake of readability, we will drop the superscript (p) or $(p\text{-}1)$ whenever the approximation order is clear from the context.

2.3. Numerical implementation

2.3.1. Solution technique

All results are computed in IEEE-754 double precision [12] using the deal.II finite element code [13]. 1) The computational mesh is obtained by globally refining a single element that covers the interval I; 2) the

Dirichlet boundary conditions are imposed strongly; 3) sufficiently accurate Gaussian quadrature formulas are used to compute the occurring integrals; 4) the UMFPACK solver [14] is used to solve the system of equations; 5) the derivatives of the numerical solution are computed in the classical finite element manner, e.g. $u_{h,x} = \sum_{i=1}^{m} u_i \varphi_{i,x}$ yields an approximation to u_x using standard FEM.

2.3.2. Error estimation

For the numerical results var_h of the variable var, the error measured in the L_2 norm is used. It is defined as

$$E_h = \|var_h - var_{\text{exc}}\|_2 \tag{10a}$$

when the exact approximation var_{exc} is available, or [15]

$$\widetilde{E_h} = \|var_h - var_{h/2}\|_2 \tag{10b}$$

otherwise, where $var_{h/2}$ is the numerical solution computed on a mesh of grid size $h/2$. Furthermore, we compute the order of convergence from either

$$Q = \log_2 \left(\frac{E_h}{E_{h/2}} \right), \tag{11a}$$

or

$$\widetilde{Q} = \log_2 \left(\frac{\widetilde{E_h}}{\widetilde{E_{h/2}}} \right), \tag{11b}$$

for which the theoretical value is one order higher than the approximation order [16].

3. Approach to finding the highest attainable accuracy

3.1. Error Evolution

The conceptual sketch of E_h against the number of DoFs (N_h) in the log-log axes can be found in Fig. 1 [17]. When N_h is relatively small, E_h may not decrease at the aforementioned theoretical order of convergence, but it basically does when N_h is relatively large. The transition from the first phase, denoted by the black circles, to the second phase, denoted by the green circles, is usually fast [17]. E_h in these two phases is controlled by the truncation error, denoted by E_{T}, and it can be represented by

$$E_h \approx E_{\text{T}} = \alpha_{\text{T}} {N_h}^{-\beta_{\text{T}}}, \tag{12}$$

in the latter phase, where α_{T} is the offset and β_{T} is the slope of the line approximating E_h. β_{T} is also the theoretical order of convergence.

When N_h is even larger, since the domination of the round-off error, denoted by E_{R}, E_h increases, see the orange circles. The slope of the line approximating E_h, denoted by β_{R}, tends to be fixed [18, 19], and its value, together with that of the offset, denoted by α_{R}, is fixed in Section 4. Thereby, E_h reading

$$E_h \approx E_{\mathrm{R}} = \alpha_{\mathrm{R}} {N_h}^{\beta_{\mathrm{R}}}, \tag{13}$$

can be predetermined in this phase.

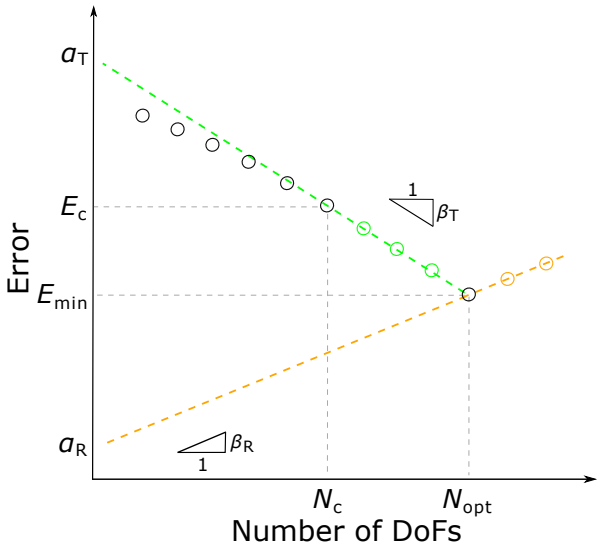


Fig. 1. Conceptual sketch of the error evolution against the number of DoFs.

3.2. Implementation process

At the beginning of the second phase, where E_h and N_h are E_{c} and N_{c}, respectively, the offset α_{T} can be inverted by using

$$\alpha_{\mathrm{T}} = E_{\mathrm{c}} / {N_{\mathrm{c}}}^{-\beta_{\mathrm{T}}}. \tag{14}$$

Thereafter, both the evolution of E_{T} and E_{R} are known. Obviously, N_{opt} occurs when $E_{\mathrm{T}} + E_{\mathrm{R}}$ is the smallest. By solving

$$\frac{d(E_{\mathrm{T}} + E_{\mathrm{R}})}{dN} = 0, \tag{15}$$

we can predict the optimal number of DoFs

$$N_{\mathrm{opt}} = \left(\frac{\alpha_{\mathrm{T}} \beta_{\mathrm{T}}}{\alpha_{\mathrm{R}} \beta_{\mathrm{R}}} \right)^{\frac{1}{\beta_{\mathrm{T}} + \beta_{\mathrm{R}}}}, \tag{16a}$$

and hence, the highest attainable accuracy

$$E_{\mathrm{min}} = \alpha_{\mathrm{T}} {N_{\mathrm{opt}}}^{-\beta_{\mathrm{T}}} + \alpha_{\mathrm{R}} {N_{\mathrm{opt}}}^{\beta_{\mathrm{R}}}. \tag{16b}$$

4. Determination of the error constants in Fig. 1

We determine α_{R} and β_{R} for the real-valued problems with only Dirichlet boundary conditions, and then validate them for problems of complex numbers and Neumann boundary conditions.

4.1. Real-valued problems with only Dirichlet boundary conditions

4.1.1. One element degree

For p=2 and p=4 for the standard FEM and mixed FEM, respectively, we consider the following equations. The Poisson equations with $u(x)$ in Table 1, for which the distribution of $\|u\|_2$ and $\|f\|_2$ is shown in Fig. 2 for Cases 1–4; the diffusion equations with $d(x)$ in Table 2 and the Helmholtz equations with $d(x)$=1 and $r(x)$ taken from the above constant $d(x)$, for which u is that of Case 1 with c=1 in Table 1.

Table 1 Settings of the Poisson equations with various $\|u\|_2$ and $\|f\|_2$.

Case	$f(x,c)$	$u(x,c)$	c
1	$-e^{-c(x-1/2)^2}\cdot$ $\left(4c^2(x-1/2)^2-2c\right)$	$e^{-c(x-1/2)^2}$	1e-4, 1e-2, 1e0, 1e2, 1e4
2	$\sin(2\pi cx)$	$(2\pi c)^{-2}\sin(2\pi cx)$	1e-2, 1e-1, 1e0, 1e1, 1e2
3	$(2\pi c)\sin(2\pi cx)$	$(2\pi c)^{-1}\sin(2\pi cx)$	
4	$\sin(2\pi cx)+1$	$(2\pi c)^{-2}\sin(2\pi cx)-\frac{x^2}{2}$	1e-4, 1e-2, 1e0, 1e2, 1e4
5	0	$c^{-1}x$	

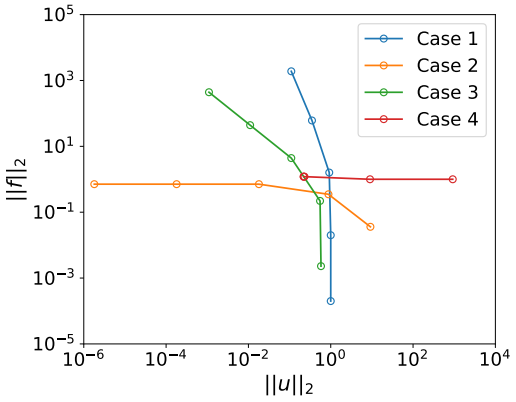
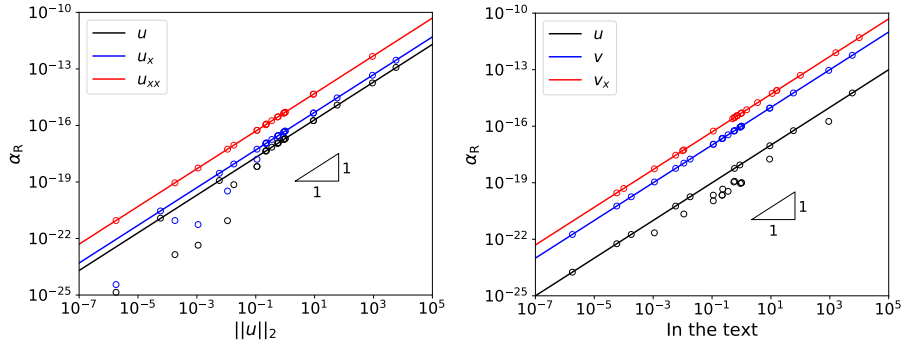


Fig. 2. Distribution of $\|u\|_2$ and $\|f\|_2$ for the Poisson equations in Table 1.

Table 2 Various $d(x)$ for the diffusion equations.

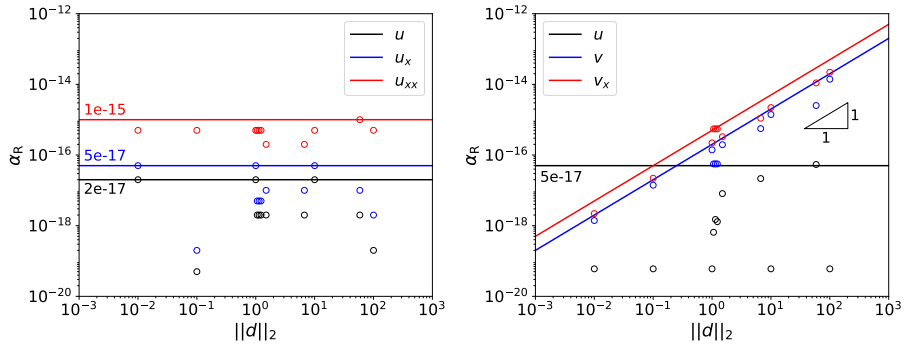
Order	$d(x)$	$\|d\|_2$	Order	$d(x)$	$\|d\|_2$
1	0.01	0.01	7	$1+\sin(10x)$	1.14
2	0.1	0.1	8	$1+\sin(100x)$	1.06
3	1	1	9	$1+x$	1.5
4	10	10	10	$1+10x$	6.7
5	100	100	11	$1+100x$	58.6
6	$1+\sin(x)$	1.23			

For all the three types of equations except for Case 5 in Table 1, for which only the round-off error exists, the error basically evolves according to that shown in Fig. 1. To summarize, β_{R} is 2 using the standard FEM and 1 using the mixed FEM. α_{R} is shown in Figs. 3 – 5, respectively. In Fig. 3(a), the ratio is 2e-17, 5e-17 and 5e-16 for u, u_x and u_{xx}, respectively. In Fig. 3(b), of which the x axis is $\|u\|_2$ for u and v, and $\|v\|_2$ for v_x, the ratio is 1e-18, 1e-16 and 5e-16 for u, v and v_x, respectively. In Fig. 4(b), the ratio is 2e-16 and 5e-16 for v and v_x, respectively. Therefore, α_{R} can be expressed as the product of a constant, for which a larger value is adopted, and an unknown that are shown in Table 3.



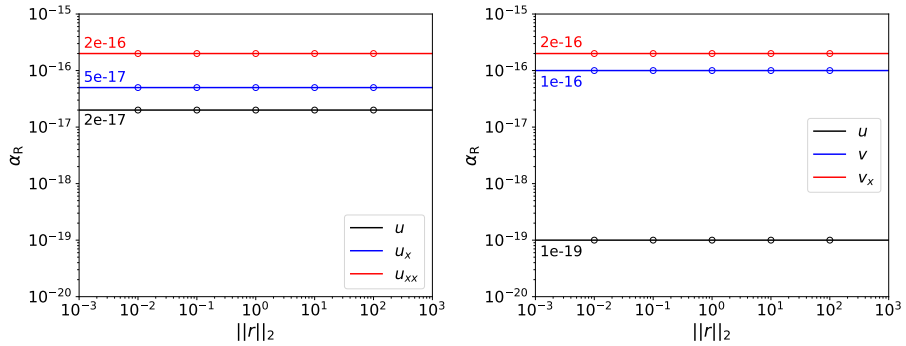
(a) The standard FEM **(b)** The mixed FEM

Fig. 3. α_{R} for the influence of $u(x)$.



(a) The standard FEM **(b)** The mixed FEM

Fig. 4. α_{R} for the influence of $d(x)$.



(a) The standard FEM **(b)** The mixed FEM

Fig. 5. α_{R} for the influence of $r(x)$.

Table 3 α_{R} in terms of the product of a constant and an unknown for the second-order differential equations.

(a) The standard FEM

	Constant	Unknown
u	2e-17	$\|u\|_2$
u_x	5e-17	
u_{xx}	1e-15	

(b) The mixed FEM

	Constant	Unknown
u	2e-17	$\|u\|_2$
v	2e-16 $\times\|d(x)\|_2$	
v_x	1e-15	$\|v\|_2$

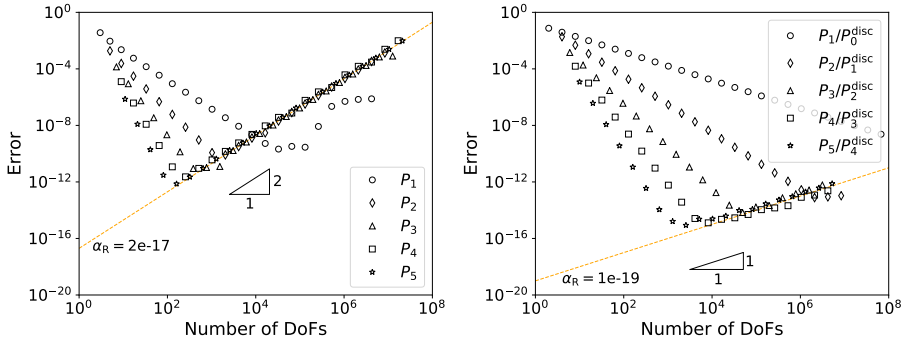
Furthermore, to make α_{R} independent of the unknowns, we propose the scaling schemes that is shown in Table 4. Note that, two schemes are required for the mixed FEM: M_1 for u and v_x, and M_2 for v. These schemes are generally able to recover the aforementioned ratios for all the Poisson equations. In what follows, the scaling scheme is used only if the scaling factor is out of [0.5, 2].

Table 4 System of equations using various scaling schemes.

	Scheme	Left-hand side	Solution	Right-hand side
The standard FEM	S	A	$\frac{1}{\|u\|_2}U$	$\frac{1}{\|u\|_2}F$
The mixed FEM	M_1	$\begin{bmatrix} M & \frac{\|u\|_2}{\|v\|_2}B \\ B^T & 0 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{\|v\|_2}V \\ \frac{1}{\|u\|_2}U \end{bmatrix}$	$\frac{1}{\|v\|_2}\begin{bmatrix} G \\ H \end{bmatrix}$
	M_2	$\begin{bmatrix} M & B \\ B^T & 0 \end{bmatrix}$	$\frac{1}{\|u\|_2}\begin{bmatrix} V \\ U \end{bmatrix}$	$\frac{1}{\|u\|_2}\begin{bmatrix} G \\ H \end{bmatrix}$

4.1.2. Multiple element degree

For p ranging from 1 to 5, we investigate the Poisson equation of Case 1 with c=1 in Table 1. An illustration of the error evolution can be found in Fig. 6 for u. α_{R} and β_{R} are basically the same for different p, and hence, higher accuracy can be obtained when using higher element degrees. Similar behaviour is also observed for the terms concerning u_x and u_{xx}, which is omitted here.



(a) The standard FEM **(b)** The mixed FEM

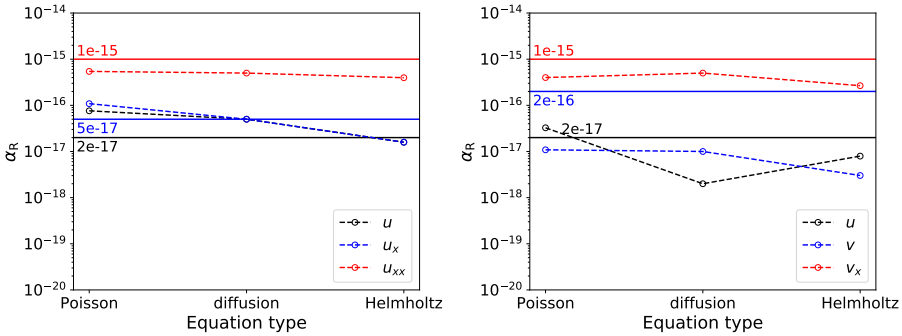
Fig. 6. An illustration of the error evolution using various p.

4.2. Problems of Neumann boundary conditions and complex numbers

We consider the equations in Table 5 for p ranging from 1 to 5, and solve them with proper scaling schemes. The error evolution resembles that in Section 4.1.2. β_{R} is only dependent on the FEM method, and α_{R} is shown in Fig. 7, in which that of v is divided by $\|d\|_2$ in Fig. 7(b). They basically fit the constants we suggest.

Table 5 Equations considering Neumann boundary conditions and complex numbers.

	Poisson	diffusion	Helmholtz
$d(x)/\|d\|_2$	$1/1$	$1+x/1.53$	$(1+i)e^{-x}/2.63$
$r(x)$	0	0	$2e^{-x}$
$f(x)$	$-e^{-(x-1/2)^2}\left(4x^2-4x-1\right)$	$-2\pi\cos(2\pi x)+$ $4\pi^2\sin(2\pi x)(x+1)$	0
Boundary conditions	$u(0)=e^{-1/4}$	$u(0)=0$	$u(0)=1$
	$u_x(1)=-e^{-1/4}$	$u_x(1)=2\pi$	$u_x(1)=0$
$\|u\|_2/\|v\|_2$	$0.92/0.5$	$0.71/4.4$	$1.26/0.75$



(a) The standard FEM **(b)** The mixed FEM

Fig. 7. α_{R} for the equations in Table 5.

5. Sensitivity analysis

Focusing on the Poisson equation in Section 4.1.2, we investigate the influence of the weak imposition of Dirichlet boundary conditions by using Weak form 3, and the iterative solver.

5.1. Weak imposition of the Dirichlet boundary condition

Using P_2 elements, the error evolution is presented in Fig. 8 for u. Since the error evolution for u_x and u_{xx} is basically the same with that using the strong imposition, we do not show them here. As can be seen, the weak imposition affects the truncation error when the penalty parameter ρ is relatively small. Thereby, we prioritize the strong imposition when using the approach in Section 3.

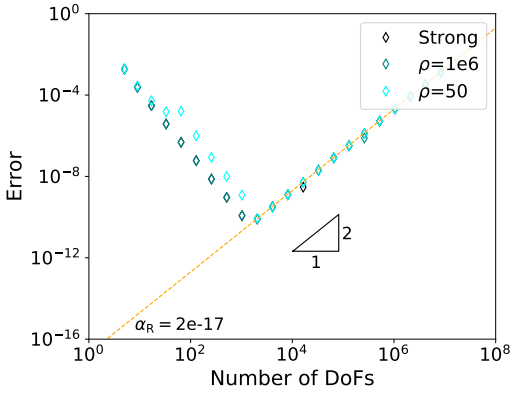


Fig. 8. Influence of the weak imposition on the error.

5.2. Solution strategy

An alternative solution method to the UMFPACK solver is the iterative Conjugate Gradient (CG) method. For the standard FEM, the CG method can be applied directly. However, for the mixed FEM, since the left-hand side of Eq. (9) is indefinite, this method can only be used after segregating Eq. (9) based on the Schur complement, which results in

$$B^\top M^{-1} B U = B^\top M^{-1} G - H, \tag{17a}$$

$$M V = G - B U. \tag{17b}$$

In the solution process of Eq. (17), the CG solver can be used for the left-hand side being either $B^\top M^{-1} B$ (Schur complement) or M. In particular, we investigate the former while the UMFPACK solver is used for the left-hand side being M.

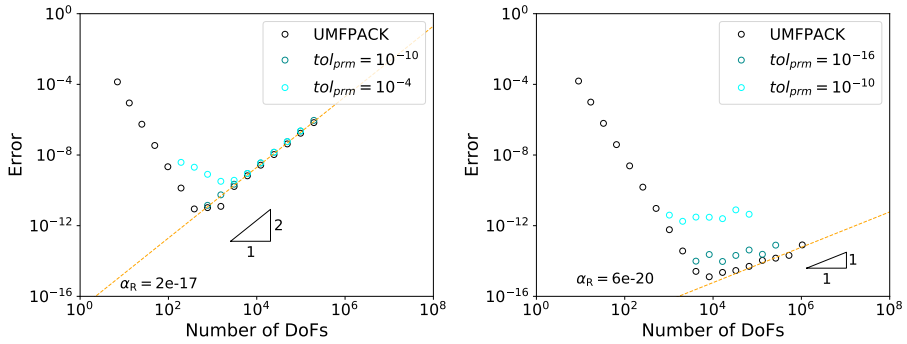
Using two tolerances, denoted by tol_{prm}, the error evolution can be found in Fig. 9 for u of the cubic approximation. In comparison with that using the UMFPACK solver, the CG solver introduces iteration errors when tol_{prm} is less strict. The accuracy using the latter is not able to be as high as that using the former for the mixed FEM. For this reason, we continue with the UMFPACK solver.

6. Algorithm and its application

Based on the approach given in section 3 and the error constants, together with the scaling schemes, provided in section 4, we introduce a practical algorithm for realizing the approach and apply it to a complex-valued Helmholtz equation.

6.1. Algorithm specifications

We define the following coefficients and use them in the steps given below.



(a) The standard FEM **(b)** The mixed FEM

Fig. 9. Influence of the iterative solver on the error.

– a minimal number of h-refinements as a precondition, denoted by $R_{\min}$, with the following default values:

$$R_{\min} = \begin{cases} 9 - p & \text{for } p < 6, \\ 4 & \text{otherwise.} \end{cases} \tag{18}$$

– the allowed maximum N_h : 10^8, denoted by $N_{\max}$.

– a stopping criterion c_s for seeking the scaling factor $\|var\|_2$ in Table 4, which is 0.001 by default.

– a relaxation coefficient c_r for seeking β_{T}, with the following default values:

$$c_r = \begin{cases} 0.9 & \text{for p} < 4, \\ 0.7 & \text{for } 4 \leqslant \text{p} < 10, \\ 0.5 & \text{otherwise.} \end{cases} \tag{19}$$

– the highest attainable accuracy determined by the computer precision, 1e-15.

Step-1. '*INPUT*'. In this step, the custom input shown in the following table has to be provided.

Table 6 Custom input of the algorithm.

Type	Item
Problem	• the differential equation to be solved
	• variables of interest
FEM	• standard or mixed formulation
	• an ordered array of element degrees $\{p_{\min}, \ldots, p_{\max}\}$

Step-2. '*NORMALIZATION*'. This step is used to find the scaling factors since they do not exist for most practical problems. The specific procedure can be found in Algorithm 1.

Algorithm 1: NORMALIZATION

while $N_h < N_{\max}$ **do**
  if $\left|\frac{\|var_h\|_2 - \|var_{2h}\|_2}{\|var_h\|_2}\right| < c_s$ **then**
    $\|var\|_2 \leftarrow \|var_h\|_2$;
    break;
  else
    $h \leftarrow h/2$;
    calculate $\|var_h\|_2$;
  end
end

Step-3. '*PREDICTION*'. The procedure for carrying out this step can be found in Algorithm 2.

Algorithm 2: PREDICTION

while $N_h < N_{\max}$ **and** $\widetilde{E_h} > E_{\mathrm{R}}$ **do**
  $\widetilde{Q} \leftarrow \log_2\left(\widetilde{E_{2h}}/\widetilde{E_h}\right)$;
  if $\widetilde{Q} \geqslant \beta_{\mathrm{T}} \times c_r$ **then**
    $N_{\mathrm{c}} \leftarrow N_h$;
    $E_{\mathrm{c}} \leftarrow \widetilde{E_h}$;
    $\alpha_{\mathrm{T}} \leftarrow E_{\mathrm{c}}/{N_{\mathrm{c}}}^{-\beta_{\mathrm{T}}}$;
    $N_{\mathrm{opt}} \leftarrow \left(\frac{\alpha_{\mathrm{T}}\beta_{\mathrm{T}}}{\alpha_{\mathrm{R}}\beta_{\mathrm{R}}}\right)^{\frac{1}{\beta_{\mathrm{R}}+\beta_{\mathrm{T}}}}$;
    $E_{\min} \leftarrow \alpha_{\mathrm{T}}{N_{\mathrm{opt}}}^{-\beta_{\mathrm{T}}} + \alpha_{\mathrm{R}}{N_{\mathrm{opt}}}^{\beta_{\mathrm{R}}}$;
  else
    $h \leftarrow h/2$;
    calculate $\widetilde{E_h}$;
  end
end

Step-4. '*OUTPUT*'. In this step, we output $E_{\min}$ obtained from *Step*-3.

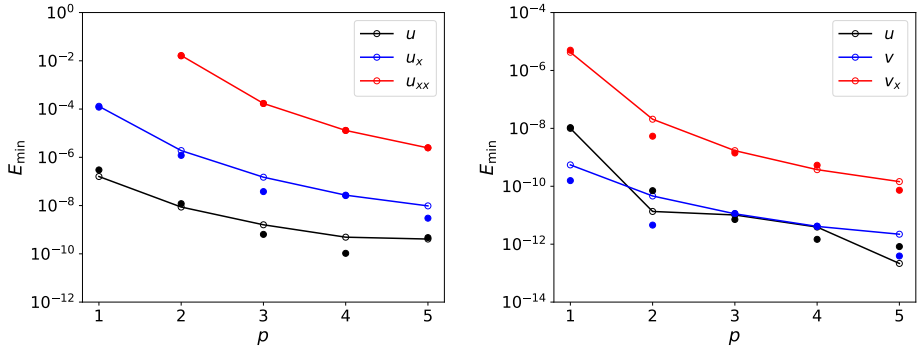
6.2. Application

We apply our algorithm to the following problem [20]:

$$((0.01+x)(1.01-x)u_x)_x - (0.01i)u(x) = 1.0, \qquad x \in I = [0,1], \tag{20}$$

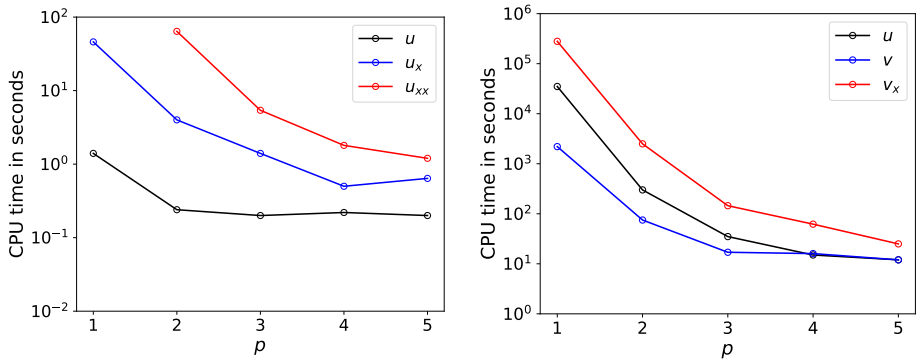
with boundary conditions $u(0)=0$ and $u_x(1)=0$. We consider the solution, and its first and second derivatives, using both the standard FEM and mixed FEM with $p \in \{1, 2, \ldots, 5\}$.

$E_{\min}$ obtained by the algorithm is shown in Fig. 10. It fits that using the brute-force approach, which uses monolithic h-refinements, well. The required CPU time is less than 0.05s and 1.00s for the standard FEM and mixed FEM, respectively. It is significantly less than that using the brute-force approach, see Fig. 11 for the latter. To compute the variable with $E_{\min}$, which is available using the brute-force approach, the CPU time is still much less. The amount of the saved CPU time in percentage is generally more than 60% and 40% for the standard FEM and mixed FEM, respectively.



(a) The standard FEM **(b)** The mixed FEM

Fig. 10. $E_{\min}$ for Eq. (20) using the algorithm. The filled circle denotes results using the brute-force approach.



(a) The standard FEM **(b)** The mixed FEM

Fig. 11. CPU time required by the brute-force approach to obtain $E_{\min}$ for Eq. (20).

7. Conclusions

A novel approach is presented to predict the highest attainable accuracy for one-dimensional second-order ordinary differential equations using the finite element methods. In contrast to the brute-force approach,

this approach uses only a few coarse grid refinements. It is viable for the solution and its first and second derivatives, using both the standard FEM and the mixed FEM with various element degrees. The algorithm for implementing the approach shows that the highest attainable accuracy can be accurately predicted while the CPU time is significantly reduced. To compute the variables with $E_{\min}$, the CPU time can be saved basically more than 60% and 40% for the standard FEM and mixed FEM, respectively. Future research will focus on the validation of the approach for two-dimensional second-order problems, where the influence of the linear-system solver, local mesh refinement and boundary conditions might be significantly different from one-dimensional problems.

Appendix A. Derivation of the weak form

Appendix A.1. The standard FEM

Multiply Eq. (1) by a test function $\eta \in H^1(I)$, and integrate it over I yield

$$\langle \eta, -(du_x)_x + ru \rangle = \langle \eta, f \rangle. \tag{A.1}$$

By applying Gauss's theorem, we obtain

$$\langle \eta_x, du_x \rangle + \langle \eta, ru \rangle = \langle \eta, f \rangle + \langle \eta, du_x n \rangle_{\Gamma_N}, \tag{A.2}$$

which gives that shown in Eq. (3). Adding auxiliary terms to the above equation renders Eq. (4).

Appendix A.2. The mixed FEM

As a first step, we introduce the auxiliary variable

$$v(x) = -d(x)u_x, \tag{A.3a}$$

allowing Eq. (1) to be rewritten as

$$-v_x - r(x)u(x) = -f(x). \tag{A.3b}$$

Multifply Eq. (A.3a) by a test function of v, i.e. $w \in H^1_{N0}(I)$, and integrate it over I yield

$$\langle d^{-1}v + u_x, w \rangle = 0. \tag{A.4a}$$

Applying Gauss's theorem to Eq. (A.4a), it becomes

$$\langle w, d^{-1}v \rangle - \langle w_x, u \rangle = -\langle w, gn \rangle_{\Gamma_D}. \tag{A.4b}$$

Multiply Eq. (A.3b) by a test function of u, i.e. $q \in L^2(I)$, and integrate it over I yield

$$-\langle q, v_x \rangle + \langle q, ru \rangle = \langle q, f \rangle. \tag{A.5}$$

Eq. (A.4b) and Eq. (A.5) result in those shown in Eq. (7).

References

[1] Mohit Kumar, Henk M. Schuttelaars, Pieter C. Roos, and Matthias Möller. Three-dimensional semi-idealized model for tidal motion in tidal estuaries. *Ocean Dynamics*, 66(1):99–118, 2016.

[2] GF Carey. Derivative calculation from finite element solutions. *Computer Methods in Applied Mechanics and Engineering*, 35(1):1–14, 1982.

[3] Joel H Ferziger and Milovan Peric. *Computational methods for fluid dynamics*. Springer Science & Business Media, 2012.

[4] Fuyun Ling and J Proakis. Numerical accuracy and stability: Two problems of adaptive estimation algorithms caused by round-off error. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, volume 9, pages 571–574. IEEE, 1984.

[5] Shan-Cong Mou, Yu-Xuan Luan, Wen-Tao Ji, Jian-Fei Zhang, and Wen-Quan Tao. An example for the effect of round-off errors on numerical heat transfer. *Numerical Heat Transfer, Part B: Fundamentals*, 72(1):21–32, 2017.

[6] Julen Alvarez-Aramberri, David Pardo, Maciej Paszynski, Nathan Collier, Lisandro Dalcin, and Victor M Calo. On round-off error for adaptive finite element methods. *Procedia Computer Science*, 9:1474–1483, 2012.

[7] Daniele Boffi, Franco Brezzi, Michel Fortin, et al. *Mixed finite element methods and applications*, volume 44. Springer, 2013.

[8] Jindrich Necas. *Direct methods in the theory of elliptic equations*. Springer Science & Business Media, 2011.

[9] Richard Haberman. *Applied partial differential equations with Fourier series and boundary value problems*. Pearson Higher Ed, 2012.

[10] Seymour Lipschutz and Marc Lipson. *Linear Algebra: Schaum's Outlines*. McGraw-Hill, 2009.

[11] Jouni Freund and Rolf Stenberg. On weakly imposed boundary conditions for second order problems. In *Proceedings of the Ninth Int. Conf. Finite Elements in Fluids*, pages 327–336. Venice, 1995.

[12] Dan Zuras, Mike Cowlishaw, Alex Aiken, Matthew Applegate, David Bailey, Steve Bass, Dileep Bhandarkar, Mahesh Bhat, David Bindel, Sylvie Boldo, et al. IEEE standard for floating-point arithmetic. *IEEE Std 754-2008*, pages 1–70, 2008.

[13] Giovanni Alzetta, Daniel Arndt, Wolfgang Bangerth, Vishal Boddu, Benjamin Brands, Denis Davydov, Rene Gassmöller, Timo Heister, Luca Heltai, Katharina Kormann, et al. The deal.II library, version 9.0. *Journal of Numerical Mathematics*, 26(4):173–183, 2018.

[14] Timothy A Davis. Algorithm 832: UMFPACK V4.3 – an unsymmetric-pattern multifrontal method. *ACM Transactions on Mathematical Software (TOMS)*, 30(2):196–199, 2004.

[15] Olof Runborg. Lecture notes in numerical solutions of differential equations (dn2255): Verifying numerical convergence rates, 2012.

[16] Mark S Gockenbach. *Understanding and implementing the finite element method*, volume 97. Siam, 2006.

[17] John Charles Butcher. *Numerical methods for ordinary differential equations*. John Wiley & Sons, 2016.

[18] Ivo Babuska and Gustaf Söderlind. On roundoff error growth in elliptic problems. *ACM Transactions on Mathematical Software*, 44(3):1–22, 2018.

[19] Meshing considerations for linear static problems. https://www.comsol.com/blogs/meshing-considerations-linear-static-problems/. Accessed: 2019-12-9.

[20] Alexander S Chernetsky, Henk M Schuttelaars, and Stefan A Talke. The effect of tidal asymmetry and temporal settling lag on sediment trapping in tidal estuaries. *Ocean Dynamics*, 60(5):1219–1241, 2010.