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 accuracy of the solution cannot increase indefinitely since the round-off error related to limited computer precision increases when the number of degrees of freedom (DoFs) is large enough. Because a priori information of the highest attainable accuracy is of great interest, we construct an innovative method to obtain the highest attainable accuracy given the order of the elements. In this method, the truncation error is extrapolated when it converges at the asymptotic rate, and the bound of the round-off error follows from a generically valid error estimate, obtained and validated through extensive numerical experiments. The highest attainable accuracy is obtained by minimizing the sum of these two types of errors. We validate this method using a one-dimensional Helmholtz equation in space. It shows that the highest attainable accuracy can be accurately predicted, and the CPU time required is much smaller compared with that using successive grid refinement.

Keywords: Finite element method, a posteriori error estimation, optimal number of degrees of freedom, *hp*-refinement strategy, round-off error.

1. Introduction

Many problems in engineering sciences and industry are modelled mathematically by initial-boundary value problems comprising systems of coupled, nonlinear partial and/or ordinary differential equations. These problems often consider complex geometries, with initial and/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 the finite difference, finite volume, and finite element methods. The latter will be adopted throughout this paper and applied to one-dimensional boundary value problems.

*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)

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

One tacitly assumes that most errors are well-balanced and/or negligibly small. In this paper, the focus is on the truncation error (E_{T}) and the round-off (E_{R}), by considering idealized problems, which do not introduce modelling errors, and using a direct solver, which avoids the introduction of iterative errors. In particular, the round-off error is often ignored based on the argument that it will be 'sufficiently small' if just IEEE-754 double-precision floating-point arithmetics [4] are adopted. Therefore, to improve the accuracy, i.e. to decrease E_{T}, one often reduces the mesh width (h-refinement), increases the approximation order (p-refinement), or applies both strategies simultaneously (hp-refinement) [5, 6].

The common characteristic of these methods is to increase the number of degrees of freedom (DoFs). However, E_{R} increases with the number of DoFs, and dominates the total error if more and more DoFs are employed [7, 8]. While typically an impractically large number of DoFs is required for E_{R} to dominate the total error if low(est)-order approximations are used, the number can be very small if high-order approximations are adopted, which are nowadays becoming more and more popular. This shift to higher order approximations makes the results more prone to be dominated by round-off errors. 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 the round-off error on the total error when using h-refinements for different orders of p. Not only the solution but also its first and second derivatives are investigated for one-dimensional, second order model problems, assuming the second derivative exists in the weak sense [9]. Both the standard finite element method (FEM) and the mixed FEM [10] are analyzed for multiple p's. Furthermore, the following factors are considered: types of boundary conditions and methods of implementing them, choices and configurations of the linear system solver, orders of magnitude of the variables and coefficients. Based on the statistics of the evolution of the round-off error, we propose an algorithm to predict the best accuracy $E_{\min}$ that occurs when the sum of E_{T} and E_{R} is the smallest, and the corresponding number of DoFs (N_{opt}).

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 illustrated in Section 3. The statistics on the evolution of the round-off error are given in Section 4. The algorithm for realizing the

approach is put forward in Section 5, followed by its validation by a Helmholtz problem 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 [11] 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 [12]:

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

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

where $f_1(x)$ and $f_2(x)$ are continuous functions defined on the unit interval I, Γ denotes the boundary of I, and x_0 denotes the values of x on 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:

$$\langle \eta_x,\, du_x\rangle + \langle \eta,\, ru\rangle = \langle \eta,\, f\rangle + \langle \eta,\, hn\rangle_{\Gamma_N} \qquad \forall \eta \in H^1_{D0}(I), \tag{3}$$

with

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

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

Imposing the Dirichlet boundary conditions in the weak sense [13], the weak form reads:

$$\boxed{\begin{array}{l}\text{Weak form 2}\\ \text{Find } u \in H^1(I) \text{ such that:}\\ \langle \eta_x,\, du_x\rangle + \langle \eta,\, ru\rangle - \langle \eta,\, du_x n\rangle_{\Gamma_D} + \langle \eta_x,\, un\rangle_{\Gamma_D} - \langle \eta,\, \rho un\rangle_{\Gamma_D}\\ \quad = \langle \eta,\, f\rangle + \langle \eta,\, hn\rangle_{\Gamma_N} + \langle \eta_x,\, gn\rangle_{\Gamma_D} - \langle \eta,\, \rho gn\rangle_{\Gamma_D} \qquad \forall \eta \in H^1(I),\\ \text{where } \rho \text{ is a positive value that serves as the penalty parameter.}\end{array}} \tag{4}$$

In both forms, η denotes the test function, n is equal to 1 at $x = 1$, and -1 at $x = 0$; the terms in the right-hand sides consist of information of Neumann boundary conditions which vanishes if no Neumann boundary conditions are 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, p is the element degree, m is the number of DoFs, which equals $p \times t + 1$, with t denoting the total number of grid cells; u_i's are the values of $u_h^{(p)}$ at the DoFs; $\varphi_i^{(p)}$'s are C^0-continuous Lagrange basis functions supported by Gauss-Lobatto points, which feature the Kronecker-delta property, i.e. $\varphi_i^{(p)}(x_j) = \delta_{ij}$, with x_j denoting the support point. This type of element will be referred to as P_p. Taking η equal to $\varphi_k^{(p)}$, $k = 1, 2, \ldots, m$, the resulting linear system of equations reads

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

where A is the stiffness matrix, F the right-hand side and $U = [u_1, \ldots, u_m]^\top$.

2.2.2. The mixed FEM

As a first step, we introduce the auxiliary variable

$$v(x) = -d(x)u_x, \tag{7a}$$

allowing Eq. (1) to be rewritten as

$$-v_x - r(x)u(x) = -f(x). \tag{7b}$$

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

Weak form 3

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

$$\langle w,\, d^{-1}v\rangle - \langle w_x,\, u\rangle = -\langle w,\, gn\rangle_{\Gamma_D} \qquad \forall w \in H^1_{N0}(I), \tag{8a}$$

$$-\langle q,\, v_x\rangle - \langle q,\, ru\rangle = -\langle q,\, f\rangle \qquad \forall q \in L^2(I), \tag{8b}$$

with

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

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

In this form, w and q denote the test functions of v and u, respectively, and n has the same interpretation as before. We approximate v and u by:

$$v \approx v_h^{(p)} = \sum_{i=1}^{m} v_i \varphi_i^{(p)}, \tag{9a}$$

$$u \approx u_h^{(p-1)} = \sum_{j=1}^{p} u_{sj} \psi_j^{(p-1)} \text{ in cell } s,\ \text{for } s = 1,\, 2,\, \ldots,\, t, \tag{9b}$$

where m is the number of DoFs for $v_h^{(p)}$, which is equal to $p \times t + 1$, v_i's are the values of $v_h^{(p)}$ at the DoFs, and $\varphi_i^{(p)}$'s are of the same type of basis functions used in Eq. (5); u_{sj}'s are the values of $u_h^{(p-1)}$ at the DoFs, $\psi_j^{(p-1)}$'s are discontinuous Lagrange basis functions, which implies that two independent u_{sj}'s have been assigned at the cell interfaces. This pair of elements will be referred to as $P_p/P_{p-1}^{\text{disc}}$. Replacing w and q by $\varphi_k^{(p)}$, $k = 1,\, 2,\, \ldots,\, p \times t + 1$, and $\psi_e^{(p-1)}$, $e = 1,\, 2,\, \ldots,\, p \times t$, respectively, the resulting coupled linear system of equations that has to be solved reads:

$$\begin{bmatrix} M & B \\ B^\top & 0 \end{bmatrix} \begin{bmatrix} V \\ U \end{bmatrix} = \begin{bmatrix} G \\ H \end{bmatrix}, \tag{10}$$

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; $V = [v_1,\, \ldots,\, v_m]^\top$ and $U = [u_{11},\, \ldots,\, u_{1p},\, \ldots,\, u_{t1},\, \ldots,\, u_{tp}]^\top$, respectively.

For the sake of readability, we will drop the superscript (p) or $(p-1)$ 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 [4] using the deal.II finite element library [14]. Unless stated otherwise, the computational mesh is obtained by globally refining a single element that covers

the interval I, and the Dirichlet boundary conditions are imposed strongly. The former means that, when the solution is real valued, the number of DoFs equals $2^R \times p+1$ using the standard FEM and $2 \times 2^R \times p+1$ using the mixed FEM, at the Rth refinement; when the solution is complex valued, the above numbers double since deal.II does not provide native support for complex-valued problems and, hence, all components need to be split into their real and imaginary parts.

To compute the occurring integrals, sufficiently accurate Gaussian quadrature formulas are used. To solve the systems of equations, the UMFPACK solver [15], which implements the multi-frontal LU factorization approach, is used unless stated otherwise. This solver results in relatively fast computations of the problems considered in this paper, and prevents the iteration errors of iterative solvers. The derivatives of the numerical solution, which are $u_{h,x}$ and $u_{h,xx}$ in the standard FEM and only $v_{h,x}$ in the mixed FEM, 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. Note that, each differentiation decreases the element degree by one.

2.3.2. Error estimation

For the numerical results var_h, where var can be u, u_x and u_{xx} of the standard FEM, and u, v and v_x of the mixed FEM, the error measured in the L_2 norm is used. It is defined as

$$E_h = \|var_h - var_{\text{exc}}\|_2 \tag{11a}$$

when the exact solution var_{exc} is available, and [16]

$$\widetilde{E_h} = \|var_h - var_{h/2}\|_2 \tag{11b}$$

otherwise, where $var_{h/2}$ is the numerical solution computed on a mesh once refined with grid size $h/2$.

2.3.3. Convergence of the solution

When the number of DoFs is relatively large, but the round-off error does not exceed the truncation error, the error converges at a fixed rate, known as asymptotic convergence rate, of which the value is one order higher than the approximation order [6]. In practice, the convergence rate in the numerical experiments can be calculated from either

$$Q = \log_2 \left(\frac{E_h}{E_{h/2}} \right) \tag{12a}$$

using Eq. (11a), or

$$\widetilde{Q} = \log_2 \left(\frac{\widetilde{E_h}}{\widetilde{E_{h/2}}} \right) \tag{12b}$$

using Eq. (11b).

3. Approach to predicting the highest attainable accuracy

A conceptual sketch of E_h against the number of DoFs (N_h) in a log-log plot can be found in Fig. 1, also see [17]. When N_h is relatively small ($N_h < N_c$), E_h does not decrease at the aforementioned asymptotic order of convergence, and only when N_h is large enough ($N_c \leqslant N_h < N_{\mathrm{opt}}$) this asymptotic order of convergence is attained. The transition from the first phase, denoted by black circles, to the second phase, denoted by green circles, is usually fast, cf. Fig. 1. E_h in both phases is controlled by the truncation error E_{T}; in the second phase E_h can be represented by

$$E_h \approx E_{\mathrm{T}} = \alpha_{\mathrm{T}} {N_h}^{-\beta_{\mathrm{T}}}. \tag{13}$$

Here α_{T} is the offset, and β_{T} is the slope of the line approximating E_h, equalling the asymptotic order of convergence, see Appendix B for the proof. Note that, α_{T} can be inverted by using

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

at the beginning of the second phase, where E_{c} equals the corresponding error E_h.

When N_h is increased too much ($N_h \geqslant N_{\mathrm{opt}}$), the round-off error E_{R} starts to dominate and E_h increases, illustrated by orange circles. At this phase, the slope of the line approximating E_h, denoted by β_{R}, tends to be fixed [8, 18]. The parameter β_{R} and the associated offset, denoted by α_{R}, are investigated in detail in Section 4. As will be shown there, α_{R} and β_{R} are fixed constants, which allows us to estimate E_h as

$$E_h \approx E_{\mathrm{R}} = \alpha_{\mathrm{R}} {N_h}^{\beta_{\mathrm{R}}}. \tag{15}$$

In summary, the evolution of E_h is described in Table 1, and depicted in Fig. 1.

Table 1 Description of the evolution of E_h.

	1. $N_h < N_{\mathrm{c}}$	2. $N_{\mathrm{c}} \leqslant N_h < N_{\mathrm{opt}}$	3. $N_{\mathrm{opt}} \leqslant N_h$
Feature	Decreasing but not converging at slope β_{T}	Decreasing and converging at slope β_{T}, with the offset α_{T}	Increasing and converging at slope β_{R}, with the offset α_{R}
Dominant error	Truncation error		Round-off error
Formula	-	$E_h \approx E_{\mathrm{T}} = \alpha_{\mathrm{T}} {N_h}^{-\beta_{\mathrm{T}}}$	$E_h \approx E_{\mathrm{R}} = \alpha_{\mathrm{R}} {N_h}^{\beta_{\mathrm{R}}}$

Since the evolution of E_h ($= E_{\mathrm{T}} + E_{\mathrm{R}}$) is known after entering the second phase, by solving

$$\frac{d(E_{\mathrm{T}} + E_{\mathrm{R}})}{dN} = 0, \tag{16}$$

we can predict the optimal number of DoFs

$$N_{\mathrm{opt}} = \left(\frac{\alpha_{\mathrm{T}} \beta_{\mathrm{T}}}{\alpha_{\mathrm{R}} \beta_{\mathrm{R}}} \right)^{\frac{1}{\beta_{\mathrm{T}} + \beta_{\mathrm{R}}}}, \tag{17a}$$

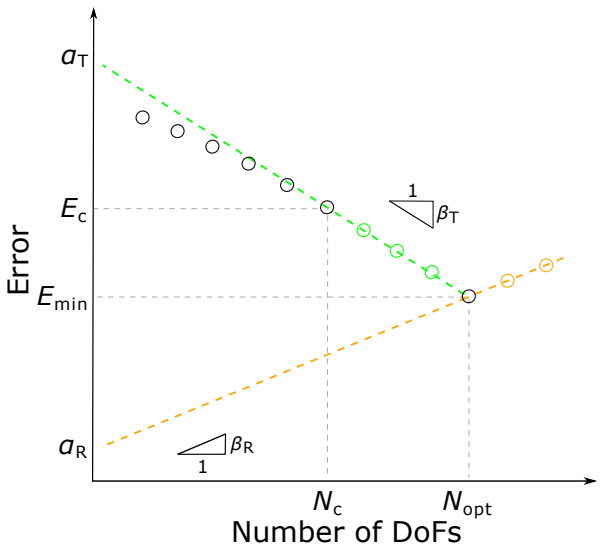


Fig. 1. Conceptual sketch of the error evolution against the number of DoFs.

and hence, the highest attainable accuracy is given by

$$E_{\min} = \alpha_{\mathrm{T}} {N_{\mathrm{opt}}}^{-\beta_{\mathrm{T}}} + \alpha_{\mathrm{R}} {N_{\mathrm{opt}}}^{\beta_{\mathrm{R}}}. \tag{17b}$$

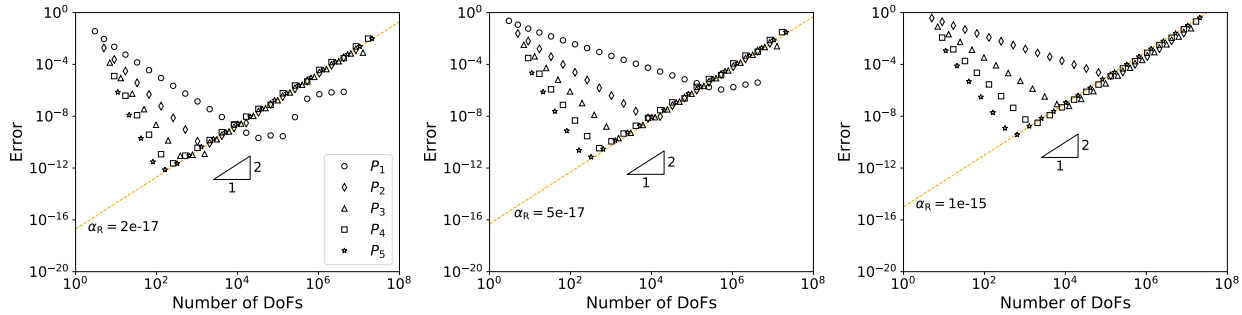
4. Results

In this section, we assess the general values of α_{R} and β_{R}. We start with a benchmark Poisson equation, for which the influences of solution strategies and boundary conditions are investigated, and then consider more general parameters for the Poisson equation, as well as the diffusion and Helmholtz equations.

4.1. Benchmark Poisson equation

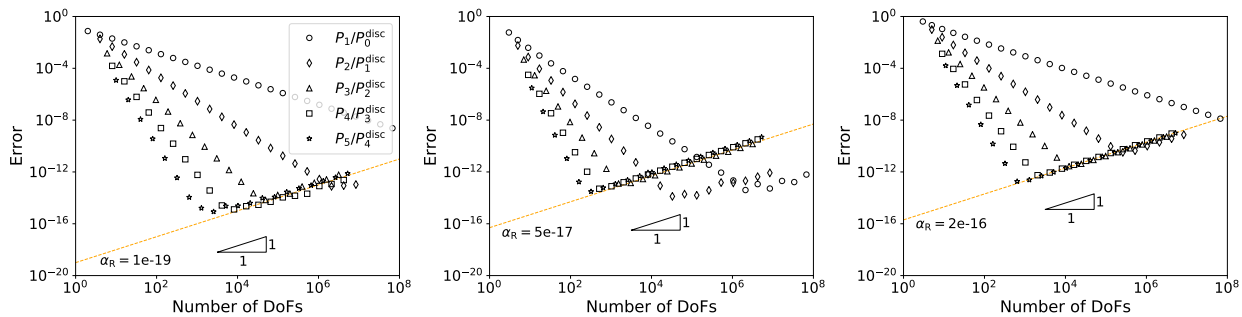
We consider the Poisson equation with $f(x) = -e^{-(x-1/2)^2}\left(4x^2 - 4x - 1\right)$. The boundary conditions are imposed as follows: $u(0) = u(1) = e^{-1/4}$. The exact solution reads $u(x) = e^{-(x-1/2)^2}$. The error E_h of the resulting solution u, and its first and second derivative, using the standard FEM and the mixed FEM, with p ranging from 1 to 5 are in Fig. 2 and Fig. 3, respectively. In these figures, α_{R} and β_{R} are denoted.

It is found that, for all dependent variables and their derivatives, the values of α_{R} and β_{R} of different element degrees are the same. The statistics of the former can be found in Fig. 4(a). The values of β_{R} only depend on the FEM method, and are 1 using the mixed FEM and 2 using the standard FEM. Notably, α_{R} is of order 10^{-16}, which is as expected when using double precision, and tends to increase slightly with increasing order of derivative. Furthermore, α_{R} of the mixed FEM is smaller than that of the standard FEM.



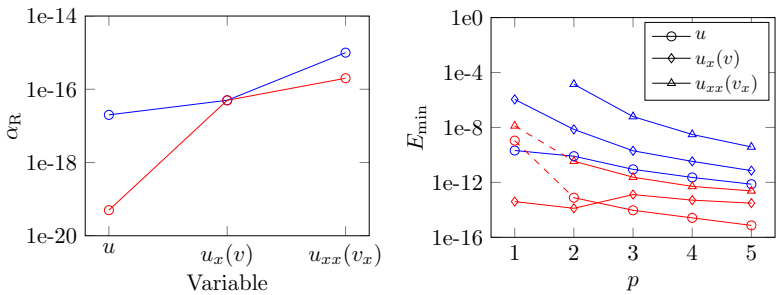
(a) u **(b)** u_x **(c)** u_{xx}

Fig. 2. Absolute errors for the benchmark Poisson equation using the standard FEM.



(a) u **(b)** v **(c)** v_x

Fig. 3. Absolute errors for the benchmark Poisson equation using the mixed FEM.



(a) α_R **(b)** E_{min}

Fig. 4. Statistics on α_R and E_{min} of the benchmark Poisson equation. The blue color denotes results using the standard FEM and the red color denotes results using the mixed FEM.

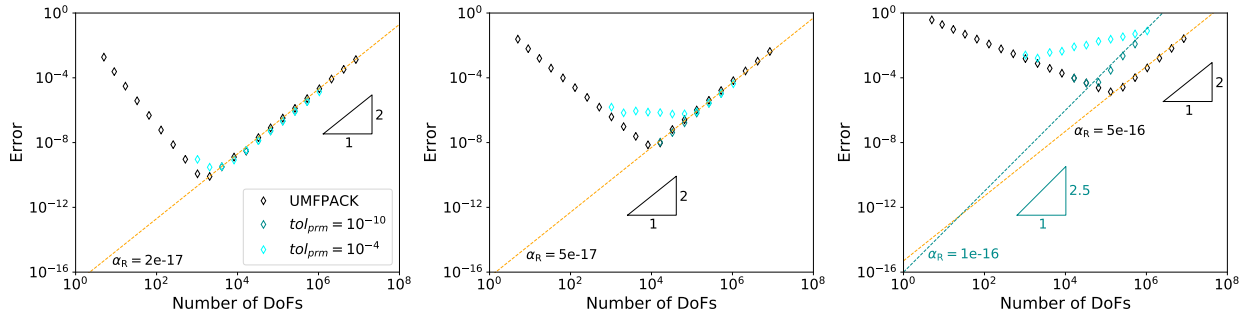
For larger p, E_T decreases faster such that smaller E_{min} can be obtained, see Fig. 4(b). In general, smaller E_{min} can be obtained using the mixed FEM compared to using the standard FEM.

In Sections 4.1.1 – 4.1.2, the sensitivity of the above results will be investigate, using P_2 elements for the standard FEM and P_4/P_3^{disc} elements for the mixed FEM.

4.1.1. Solution strategy

In this section, we investigate the influence of the solution strategy on the accuracy of the numerical solution. In particular, we compare the outcome when applying the direct solver UMFPACK with that of using the iterative Conjugate Gradient (CG) method [19], which can be applied when the left-hand side, e.g. A in Eq. (6), is symmetric and positive definite. The tolerance of the CG solver is a small number, denoted by tol_{prm}, in the standard FEM, and the product of a small number tol_{prm} and the L_2 norm of the corresponding right-hand side in the mixed FEM. When the L_2 norm of the residual, e.g. $\|F - Au\|_2$ in Eq. (6), is smaller than the tolerance, the iteration is stopped.

The standard FEM. For $tol_{prm} = 10^{-10}$ and 10^{-4}, the absolute errors of u, u_x and u_{xx} using the CG solver are shown in Fig. 5, in comparison with that using the direct solver UMFPACK.



(a) u **(b)** u_x **(c)** u_{xx}

Fig. 5. Influence of the CG solver on the accuracy using the standard FEM.

When tol_{prm} is adequately small, i.e. $tol_{prm} = 10^{-10}$, the round-off error for the solution and the first derivative using the CG solver is the same with that using the UMFPACK solver; the round-off error for the second derivative using the CG solver increases faster than that using the UMFPACK solver. When tol_{prm} is too large, i.e. $tol_{prm} = 10^{-4}$, the error contribution due to the iterative solver dominates both truncation and round-off errors for an intermediate number of DoFs.

The mixed FEM. Since the resulting matrix Eq. (10) is indefinite, a widely used alternative is to decouple the fully coupled monolithic approach using Schur's complement

$$B^\top M^{-1} BU = B^\top M^{-1} G - H, \tag{18a}$$

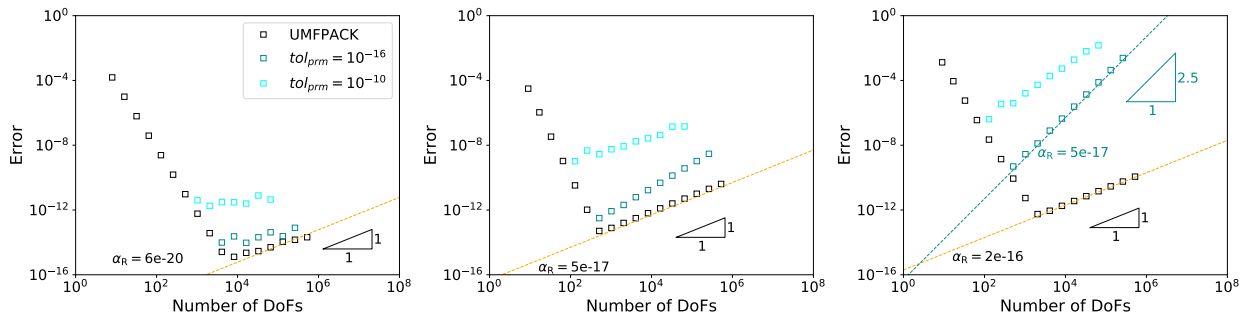
$$MV = G - BU \tag{18b}$$

and solve both equations in segregated manner, i.e. Eq. (18a) is solved in the first place to obtain U, which is then substituted into Eq. (18b) to obtain V.

Eq. (18a) involves the term $M^{-1}G$ in the right-hand side, which is computed by solving the auxiliary linear system $MY = G$ by using either the UMFPACK or the CG solver. The same options are available for solving Eq. (18b).

The difficulty in solving Eq. (18a) lies in not assembling the Schur complement matrix explicitly since it comprises M^{-1}. The CG solver only makes use of matrix-vector products of the form $(B^\top M^{-1}B)W$, which can be computed by the following three-step algorithm: $X = BW$, $MY = X$ and $Z = B^\top Y$. As before, the linear system $MY = X$ can be solved by the UMFPACK or the CG solver.

We first investigate the influence of tol_{prm} of the CG solver on the accuracy of the solutions when the left-hand side is $B^\top M^{-1}B$. In this case, the UMFPACK solver is used to solve the matrix equations when the left-hand side is M. For tol_{prm} being 10^{-16} and 10^{-10}, the results are shown in Fig. 6, in comparison with that obtained from solving the monolithic Eq. (10) directly using the UMFPACK solver. It shows that, for the problem at hand, the monolithic solution approach yields by far the most accurate solution and derivative values. The round-off error for v_x increases fastest using the Schur complement approach even though tol_{prm} is sufficiently small, i.e. $tol_{prm} = 10^{-16}$, which makes the highest attainable accuracy much lower. When tol_{prm} is less strict, i.e. $tol_{prm} = 10^{-10}$, the iteration error dominates the total error instead of the round-off error.

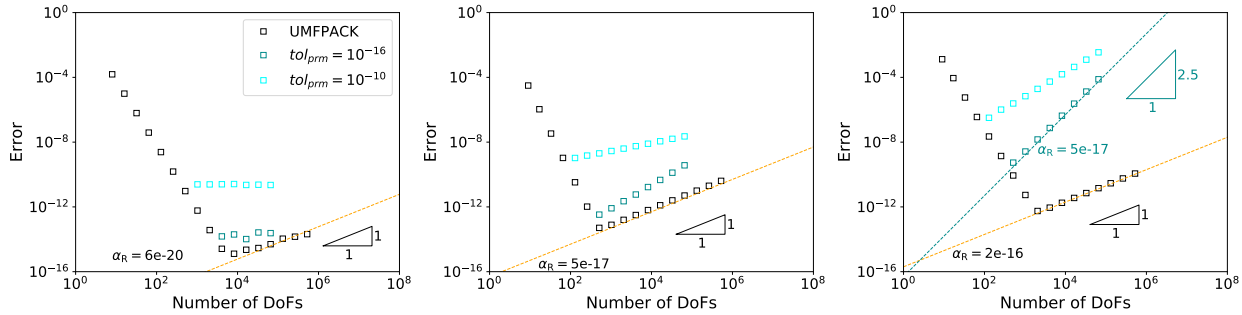


(a) u **(b)** v **(c)** v_x

Fig. 6. Influence of the CG solver on the accuracy when the left-hand side is the Schur complement using the mixed FEM.

Next, we investigate the influence of tol_{prm} of the CG solver when the left-hand side is M. In this case, the CG solver with tol_{prm} being 10^{-16} is used to solve the matrix equation with the left-hand side being $B^\top M^{-1}B$. For tol_{prm} being 10^{-16} and 10^{-10}, the results are shown in Fig. 7, in comparison with that obtained from solving the monolithic Eq. (10) directly using the UMFPACK solver. It also shows that, when the tolerance is less strict, i.e. $tol_{prm} = 10^{-10}$, the iteration error dominates the total error before the round-off error.

In summary, in comparison with the UMFPACK solver, the CG solver gives the same accuracy for u and u_x but less accuracy for u_{xx} using the standard FEM, and gives smaller accuracy for all the three variables



(a) u **(b)** v **(c)** v_x

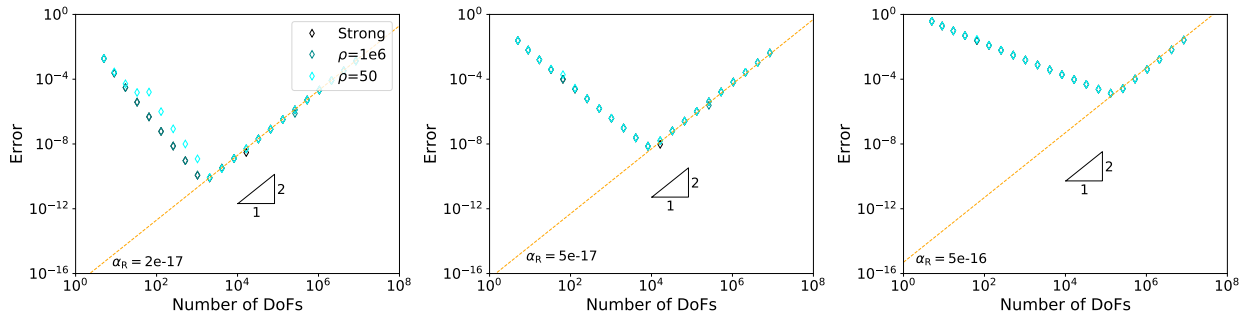
Fig. 7. Influence of the CG solver on the accuracy when the left-hand side is M using the mixed FEM.

using the mixed FEM when tol_{prm} is strict enough; the CG solver introduces iteration errors for both the standard FEM and the mixed FEM when tol_{prm} is less strict. Thereby, we continue with the UMFPACK solver in our algorithm to obtain higher accuracy.

4.1.2. Boundary condition

In this section, two aspects of the influence of the boundary conditions on the round-off error are investigated: first the method of implementing the Dirichlet boundary conditions, and secondly types of boundary conditions.

For the first aspect, using Weak form 2 for $\rho = 50$ and 10^6, the errors are depicted in Fig. 8, in comparison with that using Weak form 1. As can be seen, both the weak imposition and the strong imposition of the Dirichlet boundary condition yield the same trend line for the round-off error for the solution and its derivatives, and the magnitude of the penalty parameter in the weak imposition makes no difference. In addition, small penalty parameters might lead to larger truncation errors for u, but the difference diminishes when the penalty parameter is large enough.



(a) u **(b)** u_x **(c)** u_{xx}

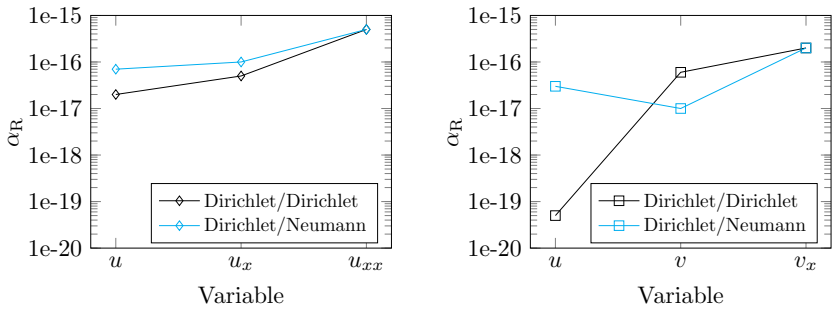
Fig. 8. Influence of the weak imposition of the Dirichlet boundary condition on the accuracy.

To construct the problem for the second aspect, the Dirichlet boundary condition at the left boundary

($x = 0$) is kept while the Dirichlet boundary condition at the right boundary ($x = 1$) is replaced by the Neumann boundary condition $u_x(1) = -e^{-1/4}$, leading to the same solution and derivative profiles.

The standard FEM. Using the standard FEM, the offsets α_{R} for the two types of boundary conditions are depicted in Fig. 9(a). For the Dirichlet/Neumann boundary condition, the offsets α_{R} for u and u_x are slightly larger than that for the Dirichlet/Dirichlet boundary condition by a factor of 3.5 and 2, respectively. The offsets α_{R} for u_{xx} are identical for the two types of boundary conditions.

The mixed FEM. Using the mixed FEM, the offsets α_{R} for the two types of boundary conditions are depicted in Fig. 9(b). As can be seen, the type of boundary conditions plays a more important role for α_{R} for the solution than α_{R} for other variables.



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

Fig. 9. Comparison of the errors for imposing Dirichlet/Dirichlet and Dirichlet/Neumann boundary conditions.

In summary, α_{R} are relatively independent of the variations in the type of boundary conditions and the method Dirichlet boundary conditions are implemented, which is an important prerequisite for our a posteriori refinement strategy to be applicable for a wide range of problems. However, since the asymptotic behaviour of the truncation error is essential in our algorithm, we continue with the strong imposition of the Dirichlet boundary condition.

In Sections 4.2 – 4.3, where the influences of $u(x)$, $d(x)$ and $r(x)$ are investigated, we only consider the Dirichlet boundary conditions, and still use P_2 elements for the standard FEM and P_4/P_3^{disc} elements for the mixed FEM.

4.2. General Poisson equation

In this section, we will again consider the Poisson equation, but now focus on the influence of the order of magnitude of the solution and right-hand side on α_{R}. To cover a wide range of scenarios, we choose the cases shown in Table 2. Each case contains a coefficient c_i, $i = 1, 2, \ldots, 5$, which is varied over several orders of magnitude so that the L_2 norm of the solution, denoted by $\|u\|_2$, and the L_2 norm of the right-hand side,

denoted by $\|f\|_2$, extend over a wide range of magnitudes. Fig. 10 gives an overview of the distribution of $\|u\|_2$ and $\|f\|_2$ for Cases 1–4.

Table 2 Setting of the Poisson equation with different right-hand sides.

Case	$f(x)$	Boundary conditions		$u(x)$
		$u(0)$	$u(1)$	
1	$\sin(2\pi c_1 x)$	0	$(2\pi c_1)^{-2}\sin(2\pi c_1)$	$(2\pi c_1)^{-2}\sin(2\pi c_1 x)$
2	$-e^{-c_2(x-1/2)^2}\cdot$ $\left(4{c_2}^2(x-1/2)^2-2c_2\right)$	$e^{-c_2/4}$	$e^{-c_2/4}$	$e^{-c_2(x-1/2)^2}$
3	$\sin(2\pi c_3 x)+1$	0	$(2\pi c_3)^{-2}\sin(2\pi c_3)-\frac{1}{2}$	$(2\pi c_3)^{-2}\sin(2\pi c_3 x)-\frac{x^2}{2}$
4	$(2\pi c_4)\sin(2\pi c_4 x)$	0	$(2\pi c_4)^{-1}\sin(2\pi c_4)$	$(2\pi c_4)^{-1}\sin(2\pi c_4 x)$
5	0	0	${c_5}^{-1}$	${c_5}^{-1}x$

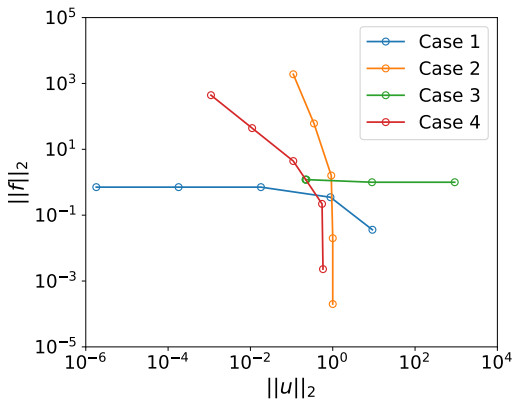
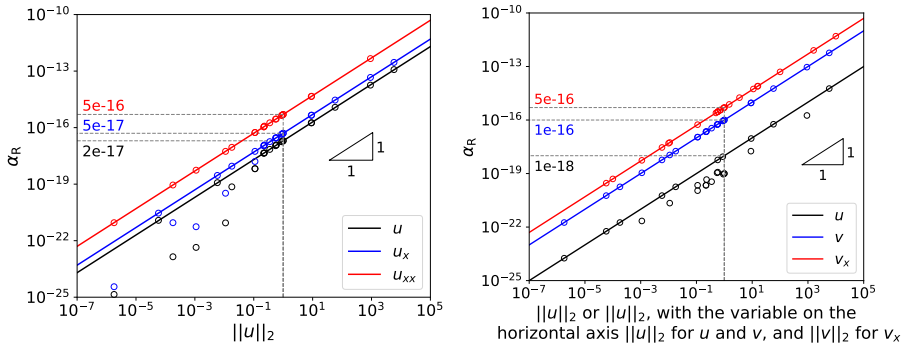


Fig. 10. Distribution of $\|u\|_2$ and $\|f\|_2$ of the Poisson equations in Table 2.

For these Poisson problems, 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} against $\|u\|_2$ is shown in Fig. 11(a) for the standard FEM; α_{R} against $\|u\|_2$ or $\|v\|_2$, depending on the dependent variable of interest, is shown in Fig. 11(b) for the mixed FEM.

For all the variables using both the standard FEM and the mixed FEM, the distribution of α_{R} can be approximated by a straight line. This implies that α_{R} has a power-law relation with the horizontal coordinate. The power term, represented by the slope of the line, is 1 for all the relations; the constant term, represented by the intercept of the line, is denoted in the figure. The latter reads 2e-17, 5e-17 and 5e-16 for u, u_x and u_{xx}, respectively, using the standard FEM, and 1e-18, 1e-16 and 5e-16 for u, v and v_x, respectively, using the mixed FEM. Note that, the intercept is the value of α_{R} when the horizontal coordinate is 1. Therefore, α_{R} has a linear relation with the L_2 norm of the dependent variable. We express α_{R} as the product of a constant and the magnitude of $\|u\|_2$ or $\|v\|_2$ that is shown in Table 3.



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

Fig. 11. α_{R} for the influence of $u(x)$.

Table 3 α_{R} in terms of the product of a constant and the L_2 norm $\|u\|_2$ or $\|v\|_2$, dependent on the specific dependent variable, for one-dimensional Poisson equations.

(a) The standard FEM

Variable	α_{R}	
u	2e-17	$\times \|u\|_2$
u_x	5e-17	
u_{xx}	5e-16	

(b) The mixed FEM

Variable	α_{R}	
u	1e-18	$\times \|u\|_2$
v	1e-16	
v_x	5e-16	$\times \|v\|_2$

4.3. General diffusion and Helmholtz equation

The coefficient $d(x)$ is taken from Table 4 for the diffusion equations, and $d(x) = 1$ is taken for the Helmholtz equations, with $r(x)$ taken from Table 5. The analytical solution u is the same as the benchmark Poisson equation, of which $\|u\|_2 = 0.92$ and $\|v\|_2 = 0.5$.

Table 4 Various $d(x)$ for the diffusion equations.

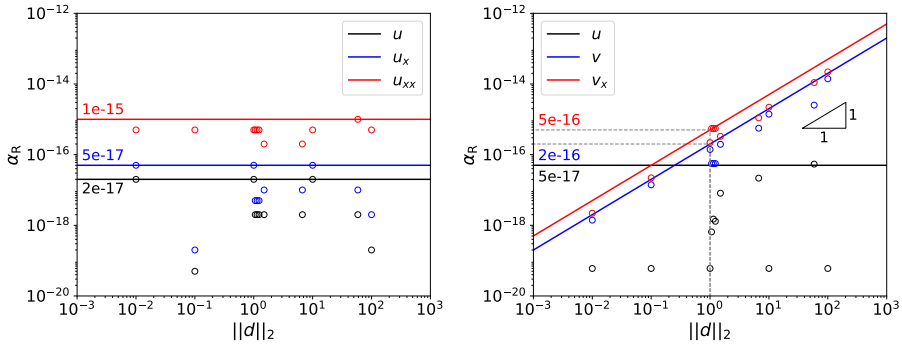
Case	$d(x)$	$\|d\|_2$	Case	$d(x)$	$\|d\|_2$
1	0.01	0.01	7	$1 + \sin(10x)$	1.14
2	0.1	0.1	8	$1 + \sin(100x)$	1.06
3	1	1	9	$1 + x$	1.5
4	10	10	10	$1 + 10x$	6.7
5	100	100	11	$1 + 100x$	58.6
6	$1 + \sin(x)$	1.23			

Again the errors show the same behaviour as that shown in Fig. 1. The coefficient β_{R} is 2 using the standard FEM and 1 using the mixed FEM.

Table 5 Various $r(x)$ for the Helmholtz equations.

Case	$r(x)$	$\|r\|_2$
1	0.01	0.01
2	0.1	0.1
3	1	1
4	10	10
5	100	100

The parameter α_{R} against $\|d\|_2$ is shown in Fig. 12 for the diffusion equations. Using the standard FEM, α_{R} is independent of $\|d\|_2$ for u, u_x and u_{xx}, with the upper bound reading 2e-17, 5e-17 and 1e-15, respectively. Using the mixed FEM, α_{R} is independent of $\|d\|_2$ for u, but is linearly dependent on $\|d\|_2$ for both v and v_x. The upper bound for u reads 5e-17; the intercept for v and v_x reads 2e-16 and 5e-16, respectively. The parameter α_{R} against $\|r\|_2$ is shown in Fig. 13 for the Helmholtz equations. Visibly, α_{R} is independent of $\|r\|_2$ using both the standard FEM and the mixed FEM. Using the standard FEM, α_{R} reads 2e-17, 5e-17 and 2e-16 for u, u_x and u_{xx}, respectively; using the mixed FEM, α_{R} reads 1e-19, 1e-16 and 2e-16 for u, v and v_x, respectively.



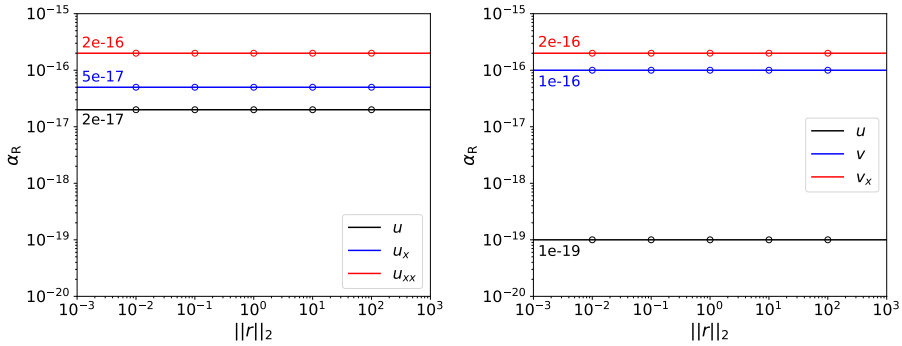
(a) The standard FEM

(b) The mixed FEM

Fig. 12. α_{R} for the general diffusion equations.

In Figs. 12–13, we do not take the influence of $\|u\|_2$ or$\|v\|_2$ into account. Since $\|u\|_2$ is of order 1, we omit its influence on α_{R} for the standard FEM and on α_{R} of u and v for the mixed FEM. Since $\|v\|_2$ is half of 1, we multiply the intercept of α_{R} of v_x in Fig. 12(b) and the upper bound of α_{R} of v_x in Fig. 13(b) by 2, for validating the constant term of α_{R} in Table 3.

Comparing the resulting α_{R} of the diffusion equations with the constant term of α_{R} in Table 3, for the standard FEM, the constant term of α_{R} of u and u_x does not change, while that of u_{xx} increases to 1e-15; for the mixed FEM, the constant term of α_{R} of u increases to 5e-17, that of v becomes 2e-16 $\times\|d\|_2$, and



(a) The standard FEM

(b) The mixed FEM

Fig. 13. α_{R} for the general Helmholtz equations.

that of v_x increases to 1e-15. Therefore, we amend α_{R} in Table 3 to be that shown in Table 6. Note that, for the constant term of α_{R} of u of the mixed FEM, considering most of its values of the diffusion equations are smaller than 2e-17, which is the value we obtain for the standard FEM, we choose 2e-17 as its value. Furthermore, since the constant term of α_{R} of the Helmholtz equations is smaller than that in Table 6, its effect can be ignored.

Table 6 α_{R} in terms of the product of a constant and the L_2 norm $\|u\|_2$ or $\|v\|_2$, dependent on the specific dependent variable, for one-dimensional second order differential equations.

(a) The standard FEM

Variable	α_{R}	
u	2e-17	$\times \|u\|_2$
u_x	5e-17	
u_{xx}	1e-15	

(b) The mixed FEM

Variable	α_{R}	
u	2e-17	$\times \|u\|_2$
v	2e-16 $\times\|d\|_2$	
v_x	1e-15	$\times\|v\|_2$

Summarizing Sections 4.2 – 4.3, the order of magnitude of the coefficient $r(x)$ is basically irrelevant; the order of magnitude of the dependent variable and the coefficient $d(x)$ can be mitigated when their values are known. Specifically, the value of the order of magnitude of the dependent variable can be obtained from a built-in function in our algorithm, and that of $d(x)$ can be obtained by carrying out simple integration.

5. A posteriori algorithm for finding the optimal number of degrees of freedom

Based on the validation experiments from the previous section, we introduce a novel a posteriori algorithm for determining $E_{\min}$ and its associated N_{opt} for the solution and its first and second derivative without performing brute-force mesh refinement. We call the algorithm *DoFinder*.

In *DoFinder*, we define the following coefficients and use them in the steps given below.

– a minimal number of h-refinements before carrying out '*NORMALIZATION*' and '*PREDICTION*', denoted by $R_{\min}$, with the following default values:

$$R_{\min} = \begin{cases} 9 - p & \text{for } p < 6, \\ 4 & \text{otherwise.} \end{cases} \tag{19}$$

We choose this parameter mainly because the error might increase, or decrease faster than the asymptotic order of convergence for coarse refinements, especially for lower-order elements.

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

– a stopping criterion c_s for seeking the L_2 norm of the dependent variable, of which the value is 0.001 by default. We choose this parameter because the analytical solution does not exist for most practical problems.

– a relaxation coefficient c_r for seeking the asymptotic order of convergence, with the following default values:

$$c_r = \begin{cases} 0.9 & \text{for } p < 4, \\ 0.7 & \text{for } 4 \leqslant p < 10, \\ 0.5 & \text{otherwise.} \end{cases} \tag{20}$$

– the offset α_{R}, see Table 6 for the default values.

The procedure of *DoFinder* consists of four steps, which are explained below:

Step-1. '*INPUT*'. In this step, the custom input shown in the Table 7 has to be provided.

Table 7 Custom input of *DoFinder*.

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*'. The function of this step is to find the L_2 norm of the dependent variable, in which elements of degree $p_{\min}$ are used. The specific procedure can be found in Algorithm 1.

Algorithm 1: NORMALIZATION

```
while N_h < N_max do
    if | (||var_h||_2 − ||var_2h||_2) / ||var_h||_2 | < c_s then
        ||var||_2 ← ||var_h||_2;
        break;
    else
        h ← h/2;
        calculate ||var_h||_2;
    end
end
```

Step-3. '*PREDICTION*'. This step finds $E_{\min}$ for each *var* and p of interest, as illustrated in Fig. 1. The procedure for carrying out this step can be found in Algorithm 2.

Algorithm 2: PREDICTION

```
while N_h < N_max and Ẽ_h > E_R do
    Q̃ ← log_2 (Ẽ_2h / Ẽ_h);
    if Q̃ ⩾ β_T × c_r then
        N_c ← N_h;
        E_c ← Ẽ_h;
        α_T ← E_c / N_c^(−β_T);
        N_opt ← ((α_T β_T)/(α_R β_R))^(1/(β_R + β_T));
        E_min ← α_T N_opt^(−β_T) + α_R N_opt^(β_R);
    else
        h ← h/2;
        calculate Ẽ_h;
    end
end
```

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

6. Validation

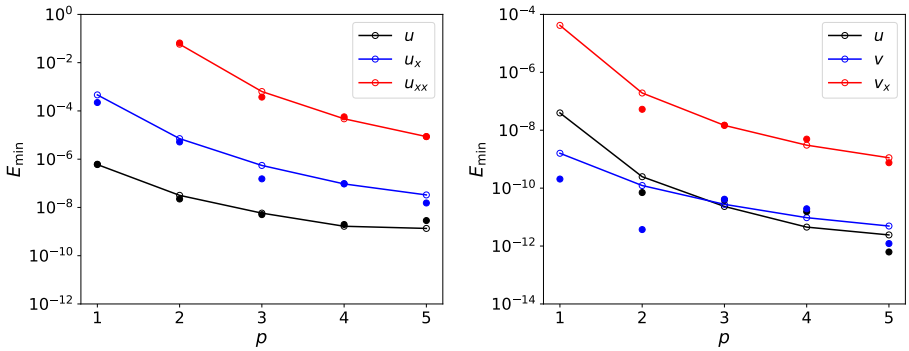
In what follows, we validate the strategy discussed in Section 3 by using the following Helmholtz problem:

$$((0.01 + x)(1.01 - x)u_x)_x - (0.01i)u(x) = 1.0, \qquad x \in I = [0, 1], \tag{21}$$

with homogeneous Dirichlet and Neumann boundary conditions imposed as follows: $u(0) = 0$ and $u_x(1) = 0$. Both the standard FEM and the mixed FEM are investigated, with the element degree p taken in $\{1, 2, \ldots, 5\}$. Variables u, u_x and u_{xx} using the standard FEM and u, v and v_x using the mixed FEM are investigated.

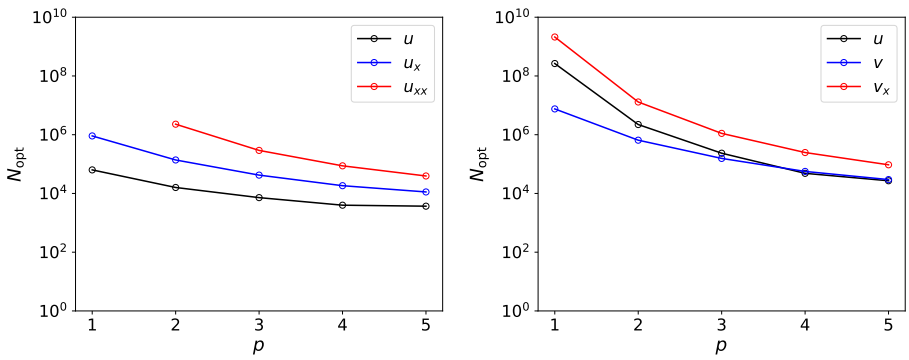
6.1. Accuracy analysis

Using *DoFinder* and the brute-force approach, $E_{\min}$'s are compared in Fig. 14. As can be seen, $E_{\min}$ can be predicted correctly. Note that, for $p = 1$ using the mixed FEM, we cannot obtain $E_{\min}$ of the brute-force approach because of the limited hardware. The associated N_{opt} using *DoFinder* is shown in Fig. 15, and it is used to compute the solution of the optimal grid in Section 6.2.



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

Fig. 14. $E_{\min}$ for Eq. (21) using *DoFinder*. The filled circle denotes results using the brute-force approach.



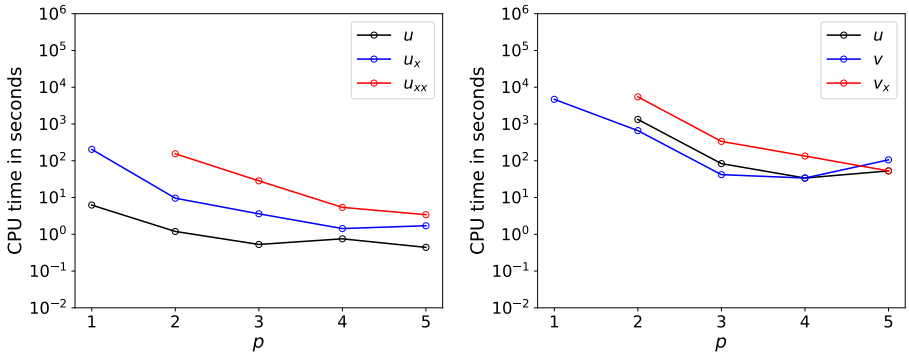
(a) The standard FEM **(b)** The mixed FEM

Fig. 15. N_{opt} for Eq. (21) using *DoFinder*.

6.2. Efficiency analysis

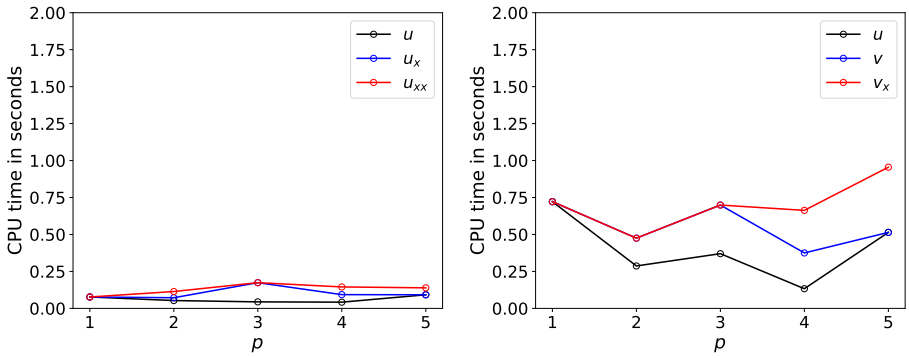
The CPU time required by *DoFinder* and the brute-force approach is shown in Fig. 16 and Fig. 17, respectively. In general, the CPU time associated with the brute-force approach decreases with increasing

element degree. In comparison with the CPU time of the brute-force approach, the CPU time of *DoFinder* is negligible.



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

Fig. 16. CPU time required by the brute-force approach to obtain $E_{\min}$ for Eq. (21).



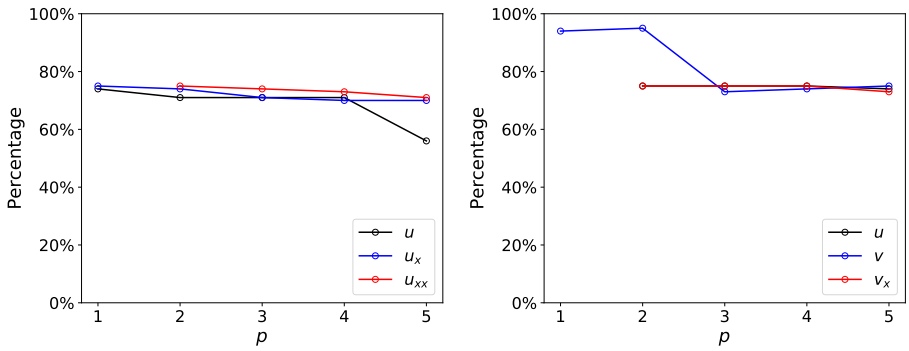
(a) The standard FEM **(b)** The mixed FEM

Fig. 17. CPU time required by *DoFinder* to obtain $E_{\min}$ for Eq. (21).

Furthermore, to obtain the solution on the optimal grid based on *DoFinder*, the computation time can be saved much in comparison with that using the brute-force approach. Fig. 18 gives the percentage of the CPU time saved by *DoFinder*. It shows a saving of the CPU time of around 70% for both the standard FEM and the mixed FEM.

6.3. Further development

For a given tolerance of a variable, denoted by tol_{var}, we are able to quickly select the available FEM method, the associated minimal available p and minimal available number of DoFs. For example, when tol_u = 1e-9, we can only achieve it using the mixed FEM, the associated minimal available p is 2, and minimal available number of DoFs is 638876.



(a) The standard FEM

(b) The mixed FEM

Fig. 18. Percentage of CPU time saved by *DoFinder* to obtain $E_{\min}$ for Eq. (21).

7. Conclusions

A novel approach is presented to predict the highest attainable accuracy for second order, ordinary differential equations using the finite element method. In contrast to the brute-force approach, which uses successive h-refinements, this approach uses only a few coarse grid refinements to reach the region of asymptotic convergence. This approach is viable for the solution and its first and second derivative, for both the standard FEM and the mixed FEM, for all element degrees. The algorithm for implementing the approach shows that the highest attainable accuracy can be accurately predicted using negligible CPU time and the CPU time to obtain this solution explicitly is significantly reduced: to compute the solution with the highest attainable accuracy using our approach results in a reduction of CPU time by around 70% for both the standard FEM and the mixed FEM. Future research will focus on the extension of this approach to 2D second order problems, where the influence of the linear system solver, local mesh refinement and boundary conditions might be significantly different from 1D 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

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

Substituting the natural boundary conditions, i.e. $d(x)u_x = h(x)$ on Γ_N, and taking $\eta = 0$ on Γ_D render Eq. (3); substituting the natural boundary conditions, not taking $\eta = 0$ on Γ_D, but adding auxiliary terms render Eq. (4).

Appendix A.2. The mixed FEM

Multiply Eq. (7a) by a test function of v, i.e. $w \in H^1_{N0}(I)$, and integrate it over I yield

$$\langle d^{-1}v + u_x, w\rangle = 0. \tag{A.3a}$$

Applying Gauss's theorem to Eq. (A.3a), it becomes

$$\langle w,\, d^{-1}v\rangle - \langle w_x,\, u\rangle = -\langle w,\, un\rangle_{\Gamma_D}. \tag{A.3b}$$

Substituting the natural boundary conditions, i.e. $u(x) = g(x)$ on Γ_D, renders Eq. (8a). Note that, unlike the standard FEM, the essential boundary conditions are imposed on Γ_N and the natural boundary conditions are imposed on Γ_D for the mixed FEM.

Multiply Eq. (7b) by a test function of u, i.e. $q \in L^2(I)$, and integrate it over I yield

$$-\langle q,\, v_x\rangle - \langle q,\, ru\rangle = -\langle q,\, f\rangle, \tag{A.4}$$

which results in Eq. (8b).

Appendix B. Proof of the slope of the decrease of the error

Here we give the proof for the standard FEM. The process for the mixed FEM is similar.

For the grid size h and element degree p, the number of DoFs

$$N_h = (1/h) \times p + 1. \tag{B.1}$$

Therefore,

$$h = \frac{p}{N_h - 1}. \tag{B.2}$$

Since the error [6]

$$E_h \leqslant Ch^{p+1}, \tag{B.3}$$

substituting Eq. (B.2) into Eq. (B.3), we obtain

$$E_h \leqslant C_1(N_h - 1)^{-(p+1)}, \tag{B.4}$$

where $C_1 = Cp^{p+1}$. Therefore, the slope is $\beta_{\mathrm{T}} = p + 1$

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] 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.

[5] B Guo and I Babuška. The hp version of the finite element method. *Computational Mechanics*, 1(1):21–41, 1986.

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

[7] 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.

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

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

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

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

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

[13] 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.

[14] 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.

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

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

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

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

[19] Theo Ginsburg. The conjugate gradient method. *Numer. Math.*, 5(1):191–200, December 1963.