Comput Mech (2007) 39: 91–111
DOI 10.1007/s00466-005-0011-z

ORIGINAL PAPER

C.A. Felippa · E. Oñate

Nodally exact Ritz discretizations of 1D diffusion–absorption and Helmholtz equations by variational FIC and modified equation methods

Received: 3 September 2005 / Accepted: 8 October 2005 / Published online: 6 January 2006

Abstract This article presents the first application of the Finite Calculus (FIC) in a Ritz-FEM variational framework. FIC provides a steplength parametrization of mesh dimensions, which is used to modify the shape functions. This approach is applied to the FEM discretization of the steady-state, one-dimensional, diffusion–absorption and Helmholtz equations. Parametrized linear shape functions are directly inserted into a FIC functional. The resulting Ritz-FIC equations are symmetric and carry a element-level free parameter coming from the function modification process. Both constant- and variable-coefficient cases are studied. It is shown that the parameter can be used to produce nodally exact solutions for the constant coefficient case. The optimal value is found by matching the finite-order modified differential equation (FOMoDE) of the Ritz-FIC equations with the original field equation. The inclusion of the Ritz-FIC models in the context of templates is examined. This inclusion shows that there is an infinite number of nodally exact models for the constant coefficient case. The ingredients of these methods (FIC, Ritz, MoDE and templates) can be extended to multiple dimensions.

Keywords Finite calculus · Variational principles · Ritz method · Functional modification · Stabilization · Finite element · Diffusion · Absorption · Helmholtz · Nodally exact solution · Modified differential equation · Templates

1 The Finite Calculus

The Finite Calculus (FIC) has been developed over the past five years [23–34,36] as a general purpose tool for improving the stability and accuracy of interior discretizations of equations of mathematical physics and engineering. Consider a problem governed by the residual equation

$$\mathbf{r}(\mathbf{u}) = \mathbf{0}, \tag{1}$$

where $\mathbf{u}$ is an array of n primary variables. These in turn are functions of the independent variables $\mathbf{x}$, which may include time. Generally Eq. (1) is an ordinary or partial differential equation, to be solved by numerical methods.

Introduce n characteristic lengths h_i collected in array $\mathbf{h}$, where each h_i is paired with the function u_i. These lengths can be viewed as linked, through as yet unspecified means, to mesh or grid dimensions. Using flux balance arguments [26,27] a modified residual is constructed

$$\mathbf{r}(\mathbf{u}) + \mathbf{r}_h(\mathbf{u}, \mathbf{h}) = \mathbf{0}. \tag{2}$$

The simplest form of $\mathbf{r}_h$ is $-(1/2)\nabla\mathbf{r}\,\mathbf{h}$, where $\nabla\mathbf{r}$ is the gradient matrix of $\mathbf{r}$ with respect to the independent variables. The discretization process, which is usually Galerkin-based FEM, is applied to Eq. 2 instead of Eq. 1. Consistency with the latter requires that $\mathbf{r}_h \to 0$ as $h_i \to 0$.

But the philosophy of FIC, as emphasized in its name, is that in practice the h_i remain *finite*. The key goal is to pick $\mathbf{r}_h$ and $\mathbf{h}$ so that stability and accuracy characteristics of the solution for *a given mesh* are improved. Further analysis of localized phenomena, such as sharp boundary layers, can be carried out by multiscale devices [6,19,30]. The FIC analysis process is diagramed in Fig. 1.

Finite Calculus has been primarily used for the solution of fluid mechanics equations involving flow, advection, diffusion, ocean waves and chemical reactions [23–34,36]. For those applications it competes with stabilization schemes such as SUPG, residual free bubbles and subgrid scale methods [3,4,9,10,19].

In a study of FIC methods for solid mechanics [37] it was found that a variational form formally analogous to the Minimum Potential Energy principle could be obtained by modifying the displacement, strain and stress fields in a manner similar to that done for the residual in the foregoing description, and adjusting their variations.

C.A. Felippa (✉)
Department of Aerospace Engineering Sciences
and Center for Aerospace Structures, University of Colorado, Boulder, Colorado 80309-0429, USA
E-mail: carlos.felippa@colorado.edu

E. Oñate
International Center for Numerical Methods in Engineering (CIMNE), Universidad Politécnica de Cataluña, Edificio C-1, c. Gran Capitán s/n 08034 Barcelona, Spain

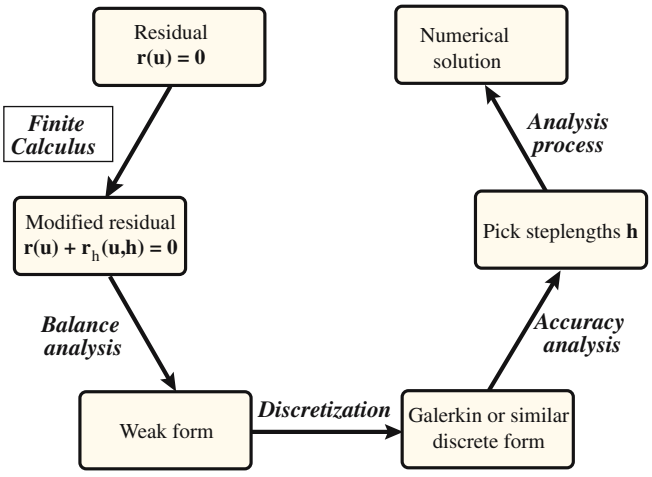

Fig. 1 The weak-form-based Finite Calculus (FIC) analysis process

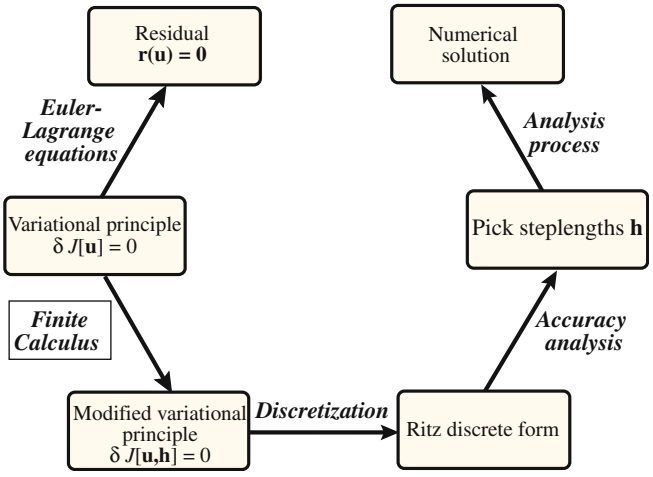

Fig. 2 The variational FIC analysis process

The approach technically falls into the class of variational principles with noncommutative variations [44], also called modified variational principles in the literature [8]. That finding provided the departure point for the present study.

2 Modified variational forms

Suppose that Eq. 1 is derivable from a functional $J[\mathbf{u}]$ in the sense that $\mathbf{r}(\mathbf{u}) = \mathbf{0}$ are the Euler-Lagrange equations of J. The first variation is

$$\delta J[\mathbf{u}] = \delta \mathbf{u}^T \, \mathbf{r}(\mathbf{u}). \tag{3}$$

Define a modified primary variable field:

$$\tilde{\mathbf{u}} \stackrel{\text{def}}{=} \mathbf{u} + \mathbf{u}_h(h), \tag{4}$$

such that $\mathbf{u}_h \to \mathbf{0}$ as $h \to 0$. The choice considered here, suggested by a previous study as noted, is

$$\tilde{\mathbf{u}} = \mathbf{u} - \tfrac{1}{2} h \boldsymbol{\alpha}^T \nabla \mathbf{u}. \tag{5}$$

Here h is an overall characteristic length, array $\boldsymbol{\alpha}$ collects scaling parameters α_i, and the factor $-1/2$ is for convenience in matching to the standard FIC method.

Substituting Eq. 5 into J yields the modified functional

$$\tilde{J}_h = J[\tilde{\mathbf{u}}] = J + J_h, \tag{6}$$

in which the augmentation term J_h vanishes as $h \to 0$. The Euler-Lagrange equation changes to

$$\delta \tilde{J}_h[\mathbf{u}] = \delta \mathbf{u}^T \left[\mathbf{r}(\mathbf{u}) + \tilde{\mathbf{r}}_h(\mathbf{u})\right]. \tag{7}$$

This has formally the same configuration as Eq. 3, and shares with it the property that as $h \to 0$ the Euler-Lagrange equation reduces to Eq. 1. But in general starting with $\mathbf{r}_h(\mathbf{u})$ of FIC, namely that in Eq. 2, does not reproduce $\tilde{\mathbf{r}}_h$. To avoid confusion we qualify Eq. 7 as the *FIC variational residual*. The functional $\tilde{J}_h$ will be called the FIC-modified functional, or FIC functional for brevity. (The superposed tildes will be eventually dropped for brevity when there is no danger of confusion.)

The numerical approximation is obtained by working with $\tilde{J}_h$ in the usual way, assuming that h is known. The residual may be used to study stability and accuracy properties of the approximation. The analysis process is diagramed in Fig. 2.

3 The modified equation method

The "Accuracy Analysis" stage of Fig. 2 is done by the *method of modified equations*. Since this is not a well known technique for differential equations, a summary along with a historical outline is presented here. An example relevant to the target application problem is worked out in Appendix A.

3.1 Backward error analysis

The conventional way to analyze accuracy of a discrete approximation is through *forward error analysis*: the amount by which the discrete solution fails to satisfy the source differential form. To make this measure practical, it is computed using local estimators such as truncation or residual errors (in FEM, through recovery from element patches). This technique furnishes *a posteriori* error indicators, and is well developed in the literature.

Backward error analysis takes the reverse approach to accuracy. Given the computed solution, it asks: which problem has the method actually solved? In other words, we seek an ODE or PDE which, *if exactly solved*, would reproduce the computed solution. This ODE or PDE is called the *modified differential equation*, often abbreviated to MoDE in the sequel. The difference between the modified equation and the original one provides an estimate of the error. An important practical advantage is that the modified equation *can be generated without actually solving the discrete problem*.

This approach is now routinely used for matrix computations after Wilkinson's definitive work in the 1960s [48–50] and has become standard part of numerical linear algebra

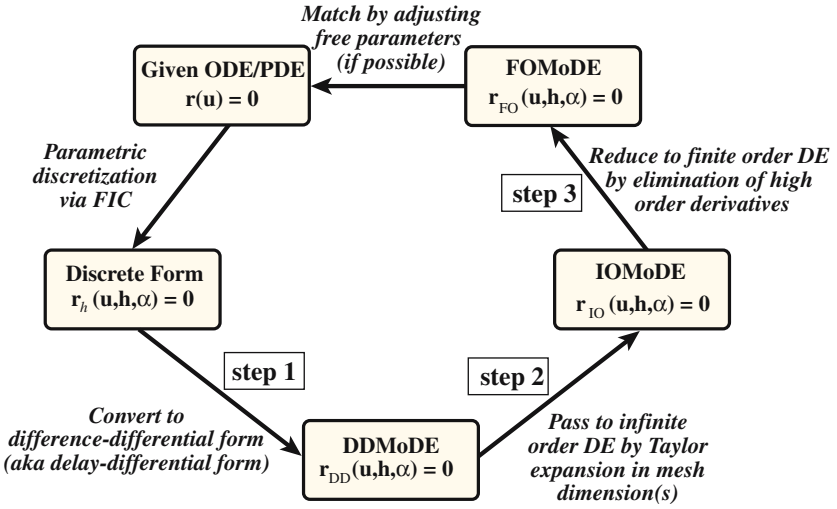

Fig. 3 Steps of the modified equation method. Achieving nodal exactness requires "closing the loop." As discussed in Sect. 3.4, this may involve additional assumptions

courses. But it is less known in differential equations. This neglect is unfortunate, since the concept follows common sense. Application problems involve physical parameters such as mass, damping, stiffness, conductivity, diffusivity, etc., which are known approximately. Transient loading actions (e.g, earthquakes, winds, waves) may be subject to high uncertainties. If the modified equation models a "nearby problem" with parameters within the range of experimental uncertainty, it is as good as the original one. This "defect correction" can be used as basis for controlling accuracy *a priori*, before any computations are actually carried out.

3.2 Applying modified equations

Let $\mathbf{r}_h(\mathbf{u}, \mathbf{h}, \boldsymbol{\alpha}) = \mathbf{0}$ denote a discretization of an ordinary or partial differential equation $\mathbf{r}(\mathbf{u}) = \mathbf{0}$. [As in Sects. 1 and 2, $\mathbf{h}$ collects lengths (in space, time or both) related to mesh or grid dimensions whereas $\boldsymbol{\alpha}$ collects free parameters]. Processing the modified equation involves three steps:

Step 1: Patch discretization → DDMoDE. The discrete equations at a typical node (a patch in FEM terminology) are rendered continuous in the independent variable(s). This produces a difference-differential form (called delay-differential form when time is the independent variable), abbreviated to DDMoDE.

Step 2: DDMoDE→ IOMoDE. The difference portion of the DDMoDE is converted to differential form by Taylor series expansion in the mesh dimensions collected in $\mathbf{h}$. This step produces a modified differential equation of infinite order, abbreviated to IOMoDE.

Step 3: IOMoDE→FOMoDE. The IOMoDE is reduced to a finite order differential equation, or FOMoDE. This is done by systematic elimination of higher order derivatives. The process typically produces an infinite series in the discretization dimensions. This series can be occasionally identified and summed in closed form. Technically this is (by far) the most difficult step. It generally requires the use of a computer algebra system (CAS) to be viable.

By comparing the FOMoDE to the original problem one can learn structural aspects of the discretization that go beyond comparison of physical parameter values. For example: preservation of Hamiltonian flow or of conservation laws in the discrete system. These are impossible or difficult to analyze with the conventional truncation error measures.

The procedural steps just outlined are flow-charted in Fig. 3. This chart also shows a parameter matching step to achieve nodal exactness, which is discussed in more detail in Sect. 3.4 below.

3.3 A brief history of modified equations

Modified differential equations as truncated forms of infinite-order ODEs appeared in conjunction with finite difference discretizations for computational fluid dynamics (CFD). The prescription for constructing them can be found in Richtmyer and Morton's textbook [40, p. 331]. Modified forms were used to interpret numerical dissipation and dispersion in the Lax-Wendroff treatment of shocks, and to derive corrective operators. Similar ideas were used by Hirt [18] and Roache [41]. A drawback of this early work is that there is no guarantee that truncation retains the relevant behavior for finite mesh dimensions, since the discarded portion could be well be dominant in coarse discretizations.

Warming and Hyett [46] were the first to describe the correct procedure for eliminating high order time derivatives of PDE space-time discretizations on the way to the finite-order modified equation (FOMoDE). (Space dimensions were treated by Fourier methods.) They attributed the

"modified equation" name to Lomax [22]. The FOMoDE forms were used for studying accuracy and stability of several CFD operators. In this work the original equation typically models flow effects of conduction and convection, **h** includes grid dimensions in space and time, and feedback is used to adjust parameters in terms of improving stability as well as reducing spurious oscillations (e.g. by artificial viscosity or upwinding) and dispersion.

The first MoDE use to study space FEM discretizations for structural mechanics can be found in [45]. However the derivative elimination and force lumping procedures were faulty, which led to incorrect conclusions. This was corrected by Park and Flaggs [38,39] who, being aware of the methods of [46], used modified equations for a systematic study of C^0 beam, plate and shell FEM discretizations.

The method has recently attracted attention from the numerical mathematics community since it provides an effective tool to understand long-time structural behavior of computational dynamic systems, both deterministic and chaotic. Recommended references are [11–14,42]. Web accessible Maple scripts for the reduction process from infinite to finite order are presented in [2]. Little of the work to date has used modified equations for optimal selection of free parameters. One exception is [12].

3.4 Free parameters for nodal exactness

Suppose that the discretization $\mathbf{r}_h(\mathbf{u}, \mathbf{h}, \boldsymbol{\alpha}) = \mathbf{0}$ contains free parameters collected in $\boldsymbol{\alpha}$. As discussed in Sects. 1,2, this is always the case for FIC discretizations, whether variationally based or not. Obviously the free parameters will carry over to the three modified equation forms: $\mathbf{r}_{\rm DD}(\mathbf{u}, \mathbf{h}, \boldsymbol{\alpha}) = \mathbf{0}$, $\mathbf{r}_{\rm IO}(\mathbf{u}, \mathbf{h}, \boldsymbol{\alpha}) = \mathbf{0}$ and $\mathbf{r}_{\rm FO}(\mathbf{u}, \mathbf{h}, \boldsymbol{\alpha}) = \mathbf{0}$. Assuming that the latter is available, the question is whether the parameters can be chosen so that

$$\mathbf{r}_{\rm FO}(\mathbf{u}, \mathbf{h}, \boldsymbol{\alpha}_{\rm M}) \equiv \mathbf{r}(\mathbf{u}), \quad \text{for any } \mathbf{h}. \tag{8}$$

Here subscript M stands for "matching." If this is possible, the discretization $\mathbf{r}_h(\mathbf{u}, \mathbf{h}, \boldsymbol{\alpha}_{\rm M})$ becomes *nodally exact*. That is, it will give the exact answer at the nodes of any discretization which generates the finite-order modified equation being matched. For FEM discretizations this scheme may be labeled a *nodally exact patch test*, since the modified equations are necessarily obtained from an element patch.

The idea is straightforward and attractive but fraught with technical difficulties. In particular:

- Exact matching may be possible only with drastic restrictions on dimensionality, system properties and discretization. For instance: constant coefficients, no source terms, regular meshes. If an exact match is impossible, some "measure of fit" (projection, minimization, etc.) has to be chosen.
- Solutions may be imaginary, non-unique, inexistent, or very hard to compute.
- The FOMoDE may contain "parasitic terms" not present in the governing equation, which cannot be cancelled out by choosing parameters. For example: the source is the Laplace equation $u_{xx} + u_{yy} = 0$ whereas the FOMoDE holds a parameter-free cross-derivative term u_{xy}. The emergence of parasitic terms was in fact observed by Park and Flaggs in their studies of C^0 plate and shell elements [38,39]. Such occurrences can be often traced to consistency defects in the discretization; in that study the presence of parasitic terms flagged element locking.
- Attaining a closed form for the FOMoDE will not be generally possible in more than one space dimension. Truncation is required. In that case the fit can at most be expected to deliver a better solution over a fixed mesh.
- Symbolic manipulations may be prohibitive, even with the help of a computer algebra system.

On the positive side, the approach is completely general, and not linked to any discretization method. The provenance of $\mathbf{r}_h(\mathbf{u}, \mathbf{h}, \boldsymbol{\alpha})$: finite elements, finite differences, boundary elements, etc., is irrelevant. It is not restricted by problem dimensionality, and does not require knowledge of exact solutions.

For FEM discretizations, the first procedure to achieve nodal exactness was Tong's adjoint technique [43]; see also [50, Appendix 7]. This requires finding exact homogeneous solutions of $\mathbf{r}(\mathbf{u}) = \mathbf{0}$, to be inserted as weight functions in a Petrov-Galerkin discretization. Related schemes are based on localized enrichment by homogeneous and/or particular solutions, for example [5,6]

3.5 Nodal exactness: advantages and limitations

Features of a nodally exact (NE) discretization are illustrated in Fig. 4. This shows the exact solution to the variable-coefficient boundary-value problem (BVP), later used in Sect. 7 as test example:

$$u'' = -750\,x\,u, \qquad u(-1) = 25,$$
$$u(1) = -4, \quad (.)'' \equiv d^2(.)/\mathrm{d}x^2. \tag{9}$$

This is compared with two FEM discretizations of six two-node elements each. That labeled NELVC comes from a

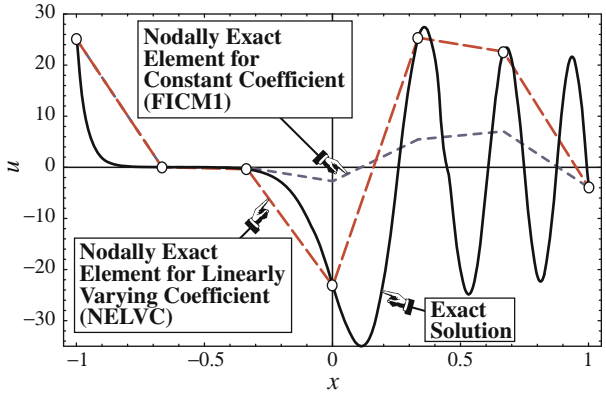

Fig. 4 Features of a nodally exact discretization for BVP Eq. 9. Solution has exponential behavior, with a sharp boundary layer, for $x < 0$. It becomes oscillatory, with increasing frequency, for $x > 0$. Node values of NELVC marked with *small circles*

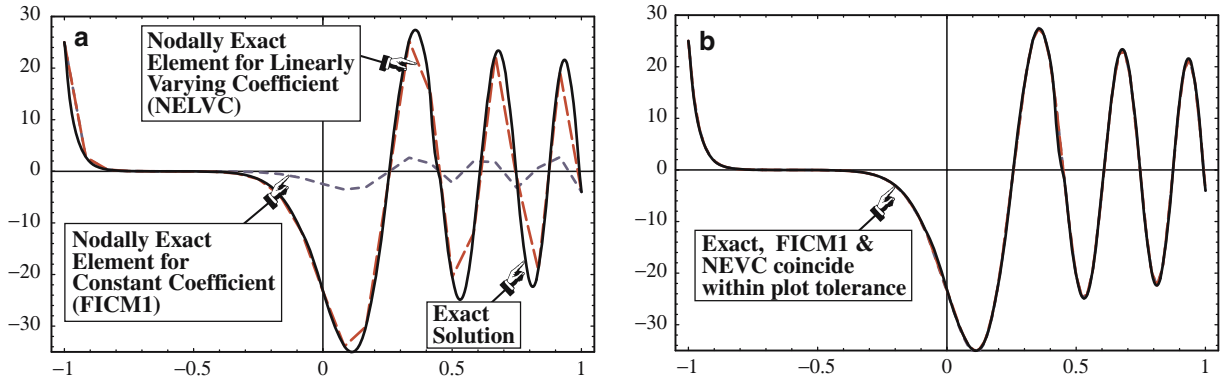

Fig. 5 Convergence behavior of discretizations for the BVP Eq. 9 : **a** 24 elements and **b** 96 elements

model that is nodally exact for a coefficient of u that varies linearly in x over each element. That labeled FICM1 is nodally exact for constant coefficients. (These two models are developed in Sects. 5 and 7, respectively, along with several others.) In both cases the FEM solution between nodes is interpolated linearly. Both approximations are plainly inadequate but for different reasons. The NELVC piecewise linear interpolation widely misses the fast variation of the exact solution, particularly on the oscillatory side.

This deficiency can be resolved by either using a better intra-element interpolation (for example, if the element was derived from shape functions) or local refinement. Advantages for multiscale modeling are nonetheless clear: an individual element can be extracted and converted into a local BVP using the end node values, and physical behavior at smaller scale introduced if appropriate. No such postprocessing is feasible with the nodally-inexact discretization, which is way off in the oscillatory side.

As the mesh is refined, the NELVC solution quickly "locks in" the correct behavior, as pictured in Fig. 5a for 24 elements. The other solution remains inadequate on the oscillatory side $x > 0$ because only three elements fit in the shorter wavelengths. Going to 96 elements the two FEM solutions cannot be distinguished from the exact one at the plotting scale of Fig. 5b. What happens is that with 12 elements per (shortest) wavelength the approximation power of the FICM1 model is finally realized, as the coefficient of $u(x)$ is sensibly constant over each element.

4 The diffusion–absorption-Helmholtz problem

Sections 4,5,6 and 7 illustrate the variational FIC discretization and the construction of modified equations for the steady-state, one-dimensional diffusion–absorption and Helmholtz equations. This problem has been recently examined by Oñate, Miquel and Hauke [36] from a FIC-Galerkin standpoint. That study includes advection terms that are not considered here. The governing differential equation that models a one-dimensional, steady state, diffusion–absorption process is

$$\frac{d}{\mathrm{d}x}\left(k\frac{\mathrm{d}u}{\mathrm{d}x}\right) - su + Q = 0, \quad \text{in} \quad x \in [x_m, x_p] \tag{10}$$

In this equation u is the state variable, $x \in [x_m, x_p]$ is the problem domain, $k \geq 0$ is the diffusion, $s \geq 0$ is the absorption (also called dissipation or destruction parameter) and Q the source term. Using primes to denote differentiation with respect to x, the foregoing ODE can be abbreviated to

$$\left(k\,u'\right)' - s\,u + Q = 0. \tag{11}$$

With the flux defined as $q = k(du/dx) = k\,u'$, the boundary conditions can be stated as

$$u = \hat{u} \quad \text{on} \quad \Gamma^u, \qquad q = \hat{q}, \quad \text{on} \quad \Gamma^q. \tag{12}$$

where Γ^u and Γ^q are the Dirichlet and Neumann boundaries, respectively. For the one-dimensional problem these consist of four combinations taken at the ends of the problem domain. This problem admits a classical variational formulation. The source functional is

$$J[u] = \int_{x_m}^{x_p} \left(\tfrac{1}{2}k(u')^2 + \tfrac{1}{2}su^2 - Qu\right)\,\mathrm{d}x. \tag{13}$$

Taking the first variation $\delta J = 0$ over admissible functions $u(x)$ that satisfy the essential BCs yields the differential Eq. 10 as Euler-Lagrange equation, and the flux constraints in Eq. 12 as natural boundary conditions.

4.1 The model problem

Following [36] and assuming $k \neq 0$, a model form of Eq. 10 is obtained by introducing the dimensionless coefficient

$$w = \frac{sa^2}{k}, \tag{14}$$

where $a = x_p - x_m$ is the length of the problem domain. This coefficient characterizes the relative importance of absorption over diffusion. The problem domain is adjusted to extend from $x_m = -1/2a$ to $x_p = 1/2a$ for convenience. We

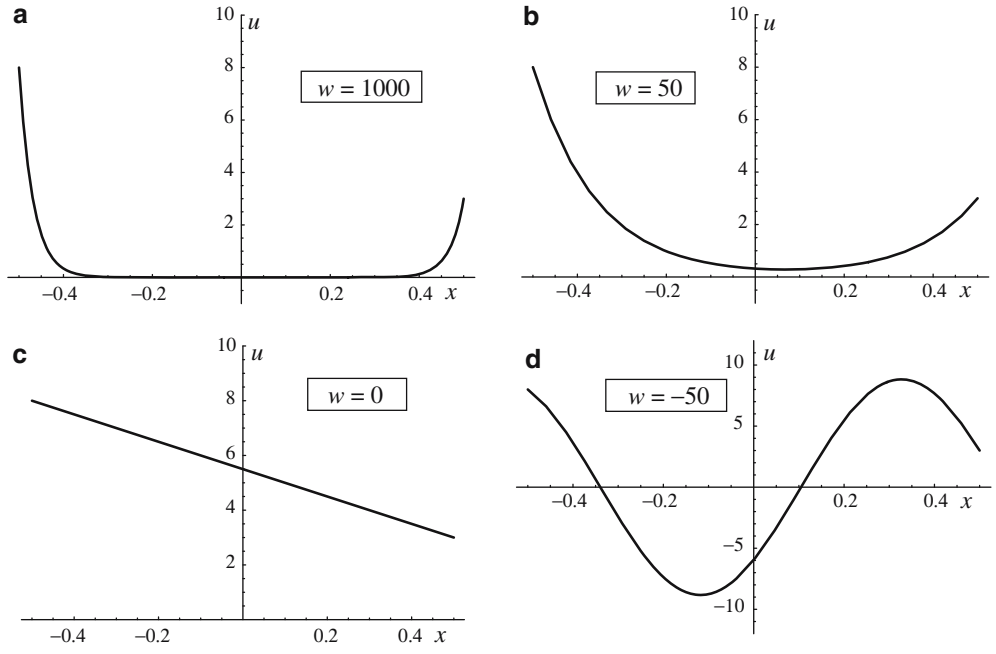

Fig. 6 Behavior of the exact solution of the model equation with $a = 1$, Dirichlet BCs $u(-1/2) = 8$ and $u(1/2) = 3$, and four values of w

assume zero source: $Q = 0$, and Dirichlet boundary conditions at both ends: $u(-1/2a) = u_m$ and $u(1/2a) = u_p$. We can now state the model problem as

$$u'' - \frac{w}{a^2}u = 0 \quad \text{for } x \in [-\tfrac{1}{2}a, \tfrac{1}{2}a],$$
$$u(-\tfrac{1}{2}a) = u_m, \quad u(\tfrac{1}{2}a) = u_p. \tag{15}$$

The associated functional is

$$J[u] = \int_{-a}^{a} \tfrac{1}{2}\left((u')^2 + \frac{w}{a^2}u^2\right) \mathrm{d}x. \tag{16}$$

where variation is taken over continuous $u(x)$ that satisfy *a priori* the Dirichlet BCs.

4.2 Exact solutions

If $w > 0$ the exact solution of the model BVP Eq. 15 can be given in term of hyperbolic functions:

$$u(x) = \frac{\sinh\left(\tfrac{1}{2}\sqrt{w}(1-\xi)\right)u_m + \sinh\left(\tfrac{1}{2}\sqrt{w}(1+\xi)\right)u_p}{\sinh(\sqrt{w})},$$
$$\xi = 2x/a. \tag{17}$$

This expression becomes $0/0$ if $w = 0$ and suffers from cancellation errors if $|w|$ is very small, say $|w| < 10^{-8}$. For that case a Taylor series about $w = 0$ gives to $O(w)$:

$$u(x) \approx \frac{1-\xi}{2}\left[1 + \frac{w}{24}(\xi^2 - 2\xi - 3)\right]u_m + \frac{1+\xi}{2}\left[1 + \frac{w}{24}(\xi^2 + 2\xi - 3)\right]u_p, \quad \xi = 2x/a. \tag{18}$$

Sample solutions are displayed in Fig. 6 for $a = 1$, $u_m = u(-1/2) = 8$, $u_p = u(1/2) = 3$, $w = 1000$, 50, 0 and -50. If $w = 0$ the solution is a straight line. As w grows, exponential boundary layers appear at Dirichlet boundaries. This is illustrated in Figs. 6a, b. If $w = 1000$ the solution is very small over most of the problem domain except for two sharp boundary layers near $x = \pm\frac{1}{2}a$.

If $w < 0$, Eq. 15 becomes the space Helmholtz equation of linear acoustics: $u'' + k^2 u = 0$ with wavenumber $k^2 = -w/a^2$. Its solution is harmonic:

$$u(x) = \frac{\sin\left(\kappa\pi(1-\xi)\right)u_m + \sin\left(\kappa\pi(1+\xi)\right)u_p}{\sin(2\kappa\pi)},$$
$$\xi = 2x/a, \quad 2\pi\,\kappa = \sqrt{-w}. \tag{19}$$

The scaled wavenumber $\kappa = \sqrt{-w}/(2\pi) = ka/(2\pi)$ represents the number of full-cycle oscillations over the domain length a. Equation 19 is only valid if $2\kappa \neq 0, 1, 2\ldots$, as otherwise the denominator vanishes. Figure 6d plots $u(x)$ for $w = -50$, in which case $\kappa = \sqrt{50}/(2\pi) = 1.1254$ cycles.

4.3 Conventional Ritz

A standard FEM solution is easily constructed by the Ritz variational formulation. Divide the domain into N^e two-node elements of length $L^e = a/N^e = \chi a$. The end nodes are i and j, with coordinates x_i and x_j, and node values u_i and u_j, respectively. Assume the piecewise linear interpolation

$$u(x) = u_i N_i(x) + u_j N_j(x), \tag{20}$$

where $N_i(x) = (x - x_i)/L^e$, $N_j = (x_j - x)/L^e$ and $L^e = x_j - x_i$ are the well known linear shape functions. Substitution

into Eq. 16 gives the element stiffness equations

$$\mathbf{S}^e\mathbf{u}^e = \frac{1}{L^e}\begin{bmatrix} 1+\frac{1}{3}\zeta & -1+\frac{1}{6}\zeta \\ -1+\frac{1}{6}\zeta & 1+\frac{1}{3}\zeta \end{bmatrix}\begin{bmatrix} u_i \\ u_j \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \chi = \frac{L^e}{a}, \quad \zeta = w\chi^2 = \frac{s(L^e)^2}{k}. \tag{21}$$

If $w = 0$ this element relation gives, upon assembly, the linear response correctly. However if $w \neq 0$, the use of Eq. 21, a scheme that may be labeled "unstabilized Ritz," displays a known defect: if w is large the solution oscillates over coarse meshes. This is illustrated in Fig. 10a for $N_e = 8$ elements and $w = 1{,}000$. Negative u values are physically incorrect if $w > 0$, which renders the solution useless.

This shortcoming is usually treated by Petrov-Galerkin stabilization schemes with suitably adjusted weight functions. The end result are unsymmetric equations for what is a self-adjoint problem.

4.4 The FIC functional

We stay within the Ritz framework and piecewise-linear shape functions, but change the functional by the method outlined in Sect. 2. For this problem, the FIC function modification technique consists of formally replacing

$$\tilde{u}(x) = u(x) - \tfrac{1}{2}hu'(x), \qquad \tilde{u}'(x) = u'(x) - \tfrac{1}{2}hu''(x). \tag{22}$$

Modified functions $\tilde{u}(x)$ and $\tilde{u}'(x)$ are inserted into Eq. 15. The tildes are then suppressed for brevity. This scheme yields a modified functional $J_h[u]$, where h is the FIC steplength. That h was derived in the original FIC by flux balancing arguments [26]. In the present context h may be simply viewed as a free parameter with dimension of length.

For piecewise linear shape functions $u''(x)$ vanishes over each element, and the second replacement in Eq. 22 may be skipped. With this simplification the modified functional is

$$J_h[u] = \int_{-a}^{a} \tfrac{1}{2}\left((u')^2 + \frac{w}{a^2}\left(u - \tfrac{1}{2}hu'\right)^2\right) \mathrm{d}x. \tag{23}$$

The Euler-Lagrange equation given by $\delta J_h[u] = 0$ is

$$(1 + \frac{wh^2}{4a^2})u'' - \frac{w}{a^2}u = 0. \tag{24}$$

From this the FIC variational residual follows as $\delta J_h = \delta u\left[(1+\frac{1}{4}wh^2/a^2)u'' - (w/a^2)u\right]$.

The Eq. (24) shows that a nonzero h injects artificial diffusion if $w > 0$. Furthermore, the sign of h makes no difference in the interior of the problem domain. As $h \to 0$ the original ODE Eq. 15 is recovered. But the key idea behind FIC is to keep h finite and directly related to mesh size.

5 The Ritz FIC equations

The FIC functional Eq. 23 is used in conjunction with the piecewise-linear interpolation Eq. 20 to construct stabilized Ritz equations for the model diffusion–absorption problem. The steplength $h = h^e$ may change from element to element. For convenience define $h^e = \alpha^e L^e$ where α^e is a dimensionless parameter to be determined over each element. The analysis that follows is restricted to constant w and elements of equal size L^e. The same α is used for all elements. This restriction is removed in Sect. 7, which studies a variable coefficient variant of the model problem.

With exact element integration (equivalently, a two-point Gauss integration rule), the following Ritz FIC element equations are obtained:

$$\mathbf{S}^e\,\mathbf{u}^e = \begin{bmatrix} S^e_{ii} & S^e_{ij} \\ S^e_{ij} & S^e_{jj} \end{bmatrix}\begin{bmatrix} u_i \\ u_j \end{bmatrix} = \frac{1}{L^e}\begin{bmatrix} 1+(\frac{1}{3}+\frac{1}{2}\alpha+\frac{1}{4}\alpha^2)\zeta^2 & -1+(\frac{1}{6}-\frac{1}{4}\alpha^2)\zeta^2 \\ -1+(\frac{1}{6}-\frac{1}{4}\alpha^2)\zeta^2 & 1+(\frac{1}{3}-\frac{1}{2}\alpha+\frac{1}{4}\alpha^2)\zeta^2 \end{bmatrix} \times \begin{bmatrix} u_i \\ u_j \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{25}$$

in which $\chi = L^e/a$ and $\zeta^2 = w\chi^2 = s(L^e)^2/k$.

5.1 Patch and modified equations

The stiffness equations for a patch of two equal-size elements comprising nodes i, j, k, as pictured in Fig. 7, are

$$\frac{1}{L^e}\begin{bmatrix} 1+(\frac{1}{3}+\frac{1}{2}\alpha+\frac{1}{4}\alpha^2)\zeta^2 & -1+(\frac{1}{6}-\frac{1}{4}\alpha^2)\zeta^2 & 0 \\ -1+(\frac{1}{6}-\frac{1}{4}\alpha^2)\zeta^2 & 2+(\frac{2}{3}+\frac{1}{2}\alpha^2)\zeta^2 & -1+(\frac{1}{6}-\frac{1}{4}\alpha^2)\zeta^2 \\ 0 & -1+(\frac{1}{6}-\frac{1}{4}\alpha^2)\zeta^2 & 1+(\frac{1}{3}-\frac{1}{2}\alpha+\frac{1}{4}\alpha^2)\zeta^2 \end{bmatrix} \times \begin{bmatrix} u_i \\ u_j \\ u_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{26}$$

To investigate choices for α we need modified equation versions of Eq. 26. The patch equation for node j is

$$S_i\,u_i + S_j\,u_j + S_k\,u_k = 0, \tag{27}$$

in which $S_i = S_k = [-12+(2-3\alpha^2)\,\zeta^2]/(12\,L^e)$ and $S_j = [24+(8+4\alpha^2)\,\zeta^2]/(12L^e)$. Equation (27) is "continuified" into a difference–differential modified equation (DDMoDE) by formally replacing $u_j \to u(x)$, $u_i \to u(x-L^e)$ and $u_k \to u(x+L^e)$ to get

$$S_i\,u(x-L^e) + S_j\,u(x) + S_k\,u(x+L^e) = 0. \tag{28}$$

Node values $u(x \pm L^e)$ are linked to $u(x)$ and its derivatives at $x = x_j$ by Taylor series:

$$u(x-L^e) = u_j - L^e u'_j + \tfrac{1}{2}(L^e)^2u''_j - \cdots,$$
$$u(x+L^e) = u_j + L^e u'_j + \tfrac{1}{2}(L^e)^2u''_j + \cdots, \tag{29}$$

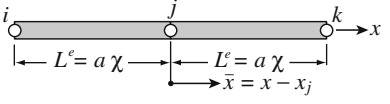

Fig. 7 A patch of two equal length elements

Replacing into Eq. 28, setting $L^e = a\chi$ and collecting terms yields the infinite-order modified equation (IOMoDE):

$$-\frac{6w}{\gamma a^2}u + \frac{1}{2!}u'' + \frac{1}{4!}a^2\chi^2 u'''' + \frac{1}{6!}a^4\chi^4 u'''''' + \cdots = 0, \tag{30}$$

in which $\gamma = 12 + (3\alpha^2 - 2)w\chi^2$. Truncating Eq. 30 to the second derivative and making $L^e \to 0$, which is the same as making $\chi \to 0$, reduces it to the original equation:

$$u'' = \frac{12w}{\gamma a^2}\,u = \frac{12w}{\left[12 + (3\alpha^2-2)w\chi^2\right]a^2}\,u \quad \stackrel{\chi=0}{\Longrightarrow} \quad u'' = \frac{w}{a^2}\,u. \tag{31}$$

The limit Eq. 31 confirms the consistency of the Ritz equations with the original ODE as $L^e \to 0$, and plays an important role in the templates developed in Sect. 6. It is not useful, however, to find nodally exact discretizations. For that all derivatives higher than two in Eq. 30 must be systematically eliminated with the techniques outlined in Appendix A. Elimination yields the finite-order modified equation (FOMoDE):

$$u'' = \frac{4}{a^2\chi^2}\left(\operatorname{arcsinh}\frac{\chi\sqrt{\mu}}{2}\right)^2 u, \quad \text{with} \quad \mu = \frac{4}{\chi^2}\left(\sinh\frac{\chi\sqrt{w}}{2}\right)^2. \tag{32}$$

Equations (26), (31) and (32) are used next to study suitable choices for parameter α.

5.2 Positivity lower bound

Suppose $u_i > 0$ and $u_k > 0$ are prescribed in the patch equations in Eq. 26. Solving for u_j from the second equation gives

$$u_j = \frac{12 + (3\alpha^2 - 2)w\chi^2}{24 + (6\alpha^2 + 8)w\chi^2}\,(u_i + u_k). \tag{33}$$

If $w \geq 0$, the denominator is positive for any $\chi \geq 0$ and real α. A non-negative u_j is guaranteed if $12+(3\alpha^2-2)w\chi^2 \geq 0$. This is met by taking $\alpha^2 \geq \alpha_P^2$, where

$$\alpha_{\rm P}^2 = \frac{2}{3} - \frac{4}{w\chi^2}. \tag{34}$$

The subscript P stands for "ensuring positivity." This result gives a useful modeling guideline: if $w\chi^2 \leq 6$, α may be set to zero, which collapses variational FIC to conventional Ritz, without impairing positivity. For example, if $w = 600$, a mesh of ten or more elements, i.e. $\chi < 1/10$, may have $\alpha = 0$ while precluding nonphysical oscillations.

Setting $\alpha^2 = \alpha_P^2$ into the element matrix $\mathbf{S}^e$ of Eq. 25 cancels out the off-diagonal terms. The assembled $\mathbf{S}$ is therefore diagonal. The solution for zero source and Dirichlet conditions at both ends is therefore zero at all interior nodes. This mimics well the boundary layer behavior for very large and positive w; say $w > 10000$. For positive but smaller w this solution can be way off, but it shows that Eq. 34 may be viewed as a *lower bound* on acceptable values of α^2, whereas the highly diffusive setting $\alpha_D^2 = 2/3$ found below is an upper bound. The results of Sect. 5.9, however, show that these bounds are of little practical value for moderate values of $w > 0$. They are also useless for the Helmholtz equation, in which case w is negative.

5.3 Diffusive upper bound

Suppose that the truncated modified equation on the left of Eq. 31 is required to reduce to the original ODE for any χ, and not just $\chi = 0$. This can be obtained by setting $\alpha^2 = \alpha_D^2$, where

$$\alpha_D^2 = \frac{2}{3}. \tag{35}$$

Computations show that using $\alpha^2 = \alpha_D^2$ overestimates the diffusion for all positive w. Thus it is "safe" in the sense of providing physically correct solutions. But these can be highly inaccurate for small or moderate w. The results of Sect. 5.9 illustrate this point. However this method gives a useful limit: if $w \to +\infty$ on a fixed mesh, $\alpha^2 \to \alpha_D^2 = 2/3$ for consistency. Consequently we have the bounds $\alpha_P^2 \leq \alpha^2 \leq \alpha_D^2$ if $w > 0$. The nodally exact value of α^2 found next lies in this interval.

5.4 Nodally exact matching

To get a nodally exact solution following the technique outlined in Sect. 3.4, require that Eq. 32 match the original equation. $u'' = (w/a^2)\,u$ for any $\{w, \chi\}$. This gives

$$\alpha_{\rm M}^2 = \frac{2}{3} - \frac{4}{w\chi^2} + \frac{1}{\sinh^2\left(1/2\chi\sqrt{w}\right)} = \frac{2}{3} - \frac{4}{\zeta^2} + \frac{1}{\sinh^2(\zeta/2)}, \tag{36}$$

where $\zeta = \chi\sqrt{w}$ and subscript M stands for matching. Equation 36 is valid for any $w > 0$. For $w \to 0$ or $\chi \to 0$ cancellations occur. These can be resolved by Taylor series expansion

$$\alpha_{\rm M}^2 = \frac{1}{3} + \frac{w\chi^2}{60} - \frac{w^2\chi^4}{1512} + \frac{w^3\chi^6}{43200} - \frac{w^4\chi^8}{1330560} + \cdots \tag{37}$$

which shows that $\alpha_{\rm M}^2 \to 1/3$ if $\chi \to 0$ or $w \to 0$. Figure 8 illustrates the variation of $\alpha_{\rm M}^2$ as function of w and χ. The latter varies between 0 and 1, attaining 1 only for one element over the domain length a. For positive w, $\alpha_{\rm M}^2$ is always positive but less than $\alpha_D^2 = 2/3$.

Extension to negative w to cover the Helmholtz equation requires some caution. A computational difficulty is that $\alpha_{\rm M}^2 > 0$ if and only if $\zeta^2 = w\chi^2 > \zeta_{\rm cut}^2$, with $\zeta_{\rm cut}^2 = -11.47463503286087328$. If $w\chi^2 < \zeta_{\rm cut}^2$, $\alpha_{\rm M}^2 < 0$ and $\alpha_{\rm M}$ is imaginary. The minimum number of elements to

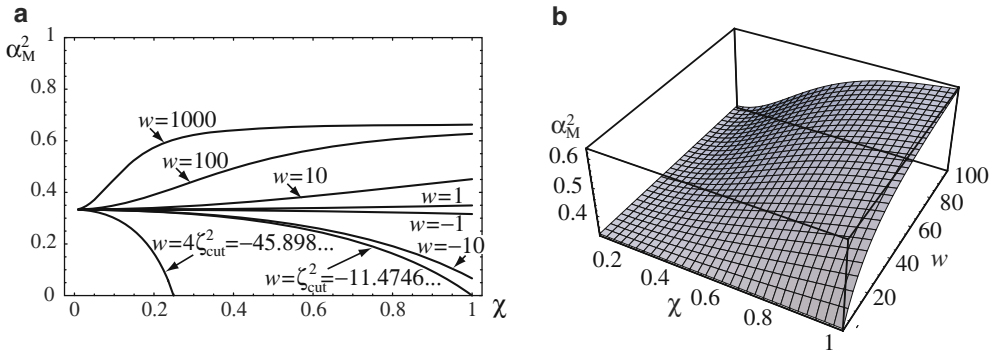

Fig. 8 Function $\alpha_{\mathrm{M}}^2(w, \chi)$. **a** 2D plots for fixed $w \in [1000, \zeta_{\mathrm{cut}}^2]$ with $\zeta_{\mathrm{cut}}^2 \approx -11.4746$. If $w\chi^2 = \zeta_{\mathrm{cut}}^2$, $\alpha_{\mathrm{M}}^2 = 0$; if $w\chi^2 < \zeta_{\mathrm{cut}}^2$, α_{M} becomes imaginary. **b** 3D plot of α_{M}^2 versus $w \in [0, 100]$ and $\chi = [0, 1]$

$$\alpha_{\mathrm{M66}}^2 = \frac{1966225060800 + 218635457040 w\chi^2 + 4430449320 w^2\chi^4 + 27010573 w^3\chi^6}{60(98311253040 + 6016210200 w\chi^2 + 115773966 w^2\chi^4 + 671585 w^3\chi^6)}. \tag{39}$$

get a positive α_{M}^2 follows from the condition $(1/N^e) = \chi \leq \sqrt{-\zeta_{\mathrm{cut}}^2}/(2\pi\kappa) = 0.539125/\kappa$, where $\kappa = \sqrt{-w}/(2\pi)$ is the number of full cycle oscillations of the exact solution over the length a, cf. Eq. 19. Solving for N^e gives

$$N^e \geq 1.85486\,\kappa, \tag{38}$$

A more easily remembered rule is $N^e \geq 2\,\kappa$, or at least two elements per wavelength. So if the solution cycles 16 times over the domain, use 32 elements or more. If Eq. 38 is not verified, however, a nodally exact solution is still obtainable but may generally require use of complex arithmetic. If the mesh consists of equal length elements and w is constant, however, complex numbers occur only in the first and last rows of the coefficient matrix, and disappear altogether on applying Dirichlet BCs.

Another peculiarity associated with the Helmholtz equation is the presence of "Dirichlet resonances." If $w = -k^2\pi^2$ for $k = 1, 2, \ldots$ then $2\pi\kappa = \sqrt{-w} = k\pi \rightarrow \sin 2\pi\kappa = 0$ and the exact solution Eq. 19 blows up. An exact number of cycles k fit over $[-a/2, a/2]$, and Dirichlet BCs with $u_m \neq u_p$ are inconsistent. In practice values of w close to those will generate large amplitude oscillations. Numerical tests in floating-point arithmetic show, however, that the nodally exact Ritz-FIC solution has no problem matching those large oscillations even if w is extremely close to a resonant value.

5.5 Pade approximants

For computer implementation, exponential functions in Eq. 36, which may cause numerical accuracy problems for $w > 10^5$, can be avoided by using Padé approximants to α_{M}^2. The (2,2), (4,4) and (6,6) diagonal approximants computed by *Mathematica* are

$$\alpha_{\mathrm{M22}}^2 = \frac{1260 + 113 w\chi^2}{30(126 + 5 w\chi^2)},$$

$$\alpha_{\mathrm{M44}}^2 = \frac{3270960 + 339948 w\chi^2 + 4787 w^2\chi^4}{189(51920 + 2800 w\chi^2 + 39 w^2\chi^4)},$$

For $\chi < \frac{1}{4}$ and $w < 10000$ these provide at least one, two and three digits of accuracy, respectively. For moderate $w\chi^2$ the higher approximants give 10–12 digits of accuracy. As $w\chi^2 \rightarrow \infty$, α_{M22}^2, α_{M44}^2, α_{M66}^2 and α_{M88}^2 approach the limits 0.75333333333333, 0.64943698277032, 0.6703190462364 and 0.66590125338669, respectively. Since α^2 should not exceed $\alpha_{\mathrm{D}}^2 = 2/3$ on account of the consistency condition discussed in Sect. 5.3, a cutoff may have to be implemented for some approximants.

5.6 Nodal exactness verification

A valuable verification of Eq. 36 can be obtained directly by equating u_j in Eq. 33 to the exact node value for a BVP posed over the two-element patch with prescribed node values u_i and u_k:

$$u_j^{\mathrm{exact}} = \tfrac{1}{2}(u_i + u_k)\ \mathrm{sech}\,\zeta = \frac{12 + (3\alpha^2 - 2)\,\zeta^2}{24 + (6\alpha^2 + 8)\,\zeta^2}\,(u_i + u_k). \tag{40}$$

Solving for α^2 and simplifying gives back Eq. 36. This method, however, cannot be used if the exact solution is not available, as it happens in two- and three-dimensional problems, whereas the modified equation method does not rely on such knowledge.

5.7 Reduced integration element

The foregoing Ritz-FIC equations have been constructed with exact element integration, which is equivalent to using a two-point Gauss rule. If a one-point Gauss reduced integration (RI) rule is used, the element equations become

$$\mathbf{S}^e\,\mathbf{u}^e = \begin{bmatrix} S_{ii}^e & S_{ij}^e \\ S_{ij}^e & S_{jj}^e \end{bmatrix} \begin{bmatrix} u_i \\ u_j \end{bmatrix} = \frac{1}{L^e} \begin{bmatrix} 1 + \frac{1}{4}(1+\alpha)^2\zeta^2 & -1 + \frac{1}{4}(1-\alpha^2)\zeta^2 \\ -1 + \frac{1}{4}(1-\alpha^2)\zeta^2 & 1 + \frac{1}{4}(1+\alpha)^2\zeta^2 \end{bmatrix} \begin{bmatrix} u_i \\ u_j \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{41}$$

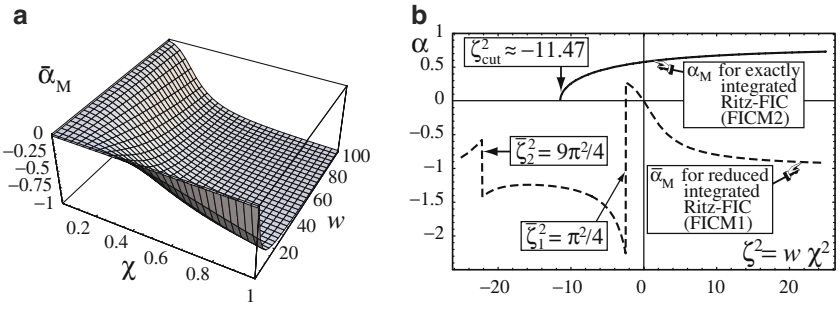

Fig. 9 Nodally exact $\bar{\alpha}_M$ for reduced-integration (RI) Ritz FIC element. **a** $\bar{\alpha}_M(\zeta, \chi)$ plotted for $w \geq 0$; **b** comparing nodally exact α_M of exactly and RI for positive and negative values of $w\chi^2$

The two-element patch equations are

$$\frac{1}{L^e}\begin{bmatrix} 1+\frac{1}{4}(1+\alpha)^2\zeta^2 & -1+\frac{1}{4}(1-\alpha^2)\zeta^2 & 0 \\ -1+\frac{1}{4}(1-\alpha^2)\zeta^2 & 2+\frac{1}{2}(1+\alpha)^2\zeta^2 & -1+\frac{1}{4}(1-\alpha^2)\zeta^2 \\ 0 & -1+\frac{1}{4}(1-\alpha^2)\zeta^2 & 1+\frac{1}{4}(1+\alpha)^2\zeta^2 \end{bmatrix} \times \begin{bmatrix} u_i \\ u_j \\ u_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{42}$$

Both α and α^2 now appear in the difference equation and the three modified equation forms. Proceeding as before one obtains the nodally exact α as

$$\bar{\alpha}_M = \frac{1-(\tau/\zeta)\sqrt{\zeta^2-4+8/\tau-4/\tau^2}}{\tau-1}, \qquad \zeta = \chi\sqrt{w}, \quad \tau = \cosh\zeta. \tag{43}$$

The variation of $\bar{\alpha}_M(w, \chi)$ is plotted in Fig. 9a. Its Taylor series is

$$\bar{\alpha}_M = -\frac{1}{6}w\chi^2 - \frac{13}{720}w^2\chi^4 + \frac{127}{20160}w^3\chi^6 + \cdots \tag{44}$$

which shows that $\bar{\alpha}_M \to 0$ as $w \to 0$ or $\chi \to 0$. The variation of $\bar{\alpha}_M(w, \chi)$ for the diffusion–absorption case $w > 0$ is plotted in Fig. 9a. It is negative and smooth.

The behavior of $\bar{\alpha}_M$ for the Helmholtz equation is more complicated than that of α_M, the positive square root of Eq. 36. Both are plotted as functions of $\zeta^2 = w\chi^2$ for $\zeta^2 \in [-25, 25]$ in Fig. 9b. Whereas α_M is real for $w\chi^2 \geq \zeta^2_{cut} = -11.474635\ldots$ and varies smoothly, $\bar{\alpha}_M$ is real for all ζ^2 but jumps at $\bar{\zeta}^2_k = -k^2\pi^2/4$ for $k = 1, 2, \ldots$. These are roots of $\cosh(i|\bar{\zeta}_k|) = \cos|\bar{\zeta}_k| = 0$. These jumps are harmless, however, since they occur at the set of 'Dirichlet resonances" discussed in Sect. 4.2, where the exact solution Eq. 19 blows up. In fact the RI model has the advantage that $\bar{\alpha}_M$ stays real for any w, whether positive or negative.

Using the exact solution to find $\bar{\alpha}_M$ gives two solutions: $1 \pm (\tau/\zeta)\sqrt{\zeta^2-4+8/\tau-4/\tau^2}/(\tau-1)$ with $\tau = \cosh\zeta$, which appear as roots of a quadratic. Taking the minus sign reproduces Eq. 43, whereas taking the plus sign gives an $\bar{\alpha}_M$ that "blows up" if $w\chi^2 \to 0$, and is therefore less desirable.

5.8 Source terms

The MoDE treatment of a smooth source term $q(x)$ in $u'' - (w/a^2)u = q(x)$ can be done by entirely analogous techniques, but there are representation choices. One is based on expanding $q(x)$ in Taylor series: $q(x) = q(x_j) + q'(x_j)(x - x_j)/a + \cdots$ at node j, and inserting into the FIC functional to derive a consistent node force term q_j by the usual methods. Then $q(x_j), q'(x_j)\ldots$ appear in the RHS of the IOMoDE→FOMode elimination system, and the solution series is identified into the FOMoDE. Alternatively $q(x)$ can be expanded in Fourier series [38,39]. Delta function source terms can be processed directly.

5.9 Numerical results

This section present numerical results obtained with the Ritz-FIC method for the model problem Eq. 15. The problem domain is taken to have unit length ($a = 1$) extending from $x_m = -1/2a = -1/2$ through $x_p = 1/2a = 1/2$. The boundary conditions are of Dirichlet type: $u(-1/2) = 8$ and $u(1/2) = 3$. The domain is divided into eight elements of equal size; thus $\chi = L^e/a = 1/8$. Four values of w: 1000, 50, −50 and −1,000, are tested. Results are plotted in Fig. 10, and listed to ten decimal places in Tables 1, 2, 3 and 4. Labels shown in those tables are template names assigned in Sect. 6.

Results for $w = 1000$. The solution exhibits two sharp boundary layers. Over the propagation region, which extends roughly over the middle six elements of this discretization, $u(x)$ takes small positive values, of order 10^{-3} or less. The problem is discretized using four choices of α: $\alpha = 0$ (conventional Ritz), $\alpha^2_C = 2/3$, $\alpha^2_P = 101/375 = 0.410667$ and $\alpha^2_M = 0.490503$. Numerical results are shown in Fig. 10a, and listed in Table 1. As expected the solution for α^2_M is nodally exact. The results for $\alpha = 0$ oscillate giving unacceptable negative values. Results for α_D and α_P give the correct physical behavior, and bound the boundary layer behavior on both sides. Although the difference of results computed for α_D and α_P with the exact solution are masked in the scale of the plot, discrepancies at interior points are clear from Table 1.

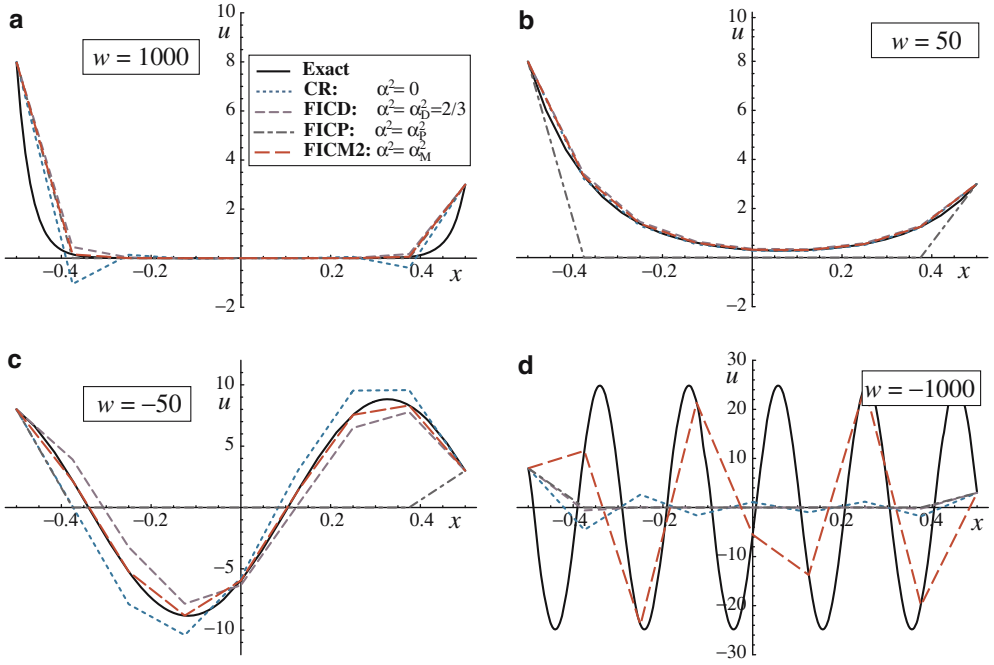

Fig. 10 Ritz-FIC results for eight-element discretization of the diffusion–absorption model problem with: **a** $w = 1000$, **b** $w = 50$, **c** $w = -50$, **d** $w = -1000$. Dirichlet BCs $u(-\frac{1}{2}) = 8$ and $u(\frac{1}{2}) = 3$, for four choices of α, compared to the exact solution

Table 1 Ritz-FIC eight-element solutions, $w = 1000$

Node	Exact	$\alpha^2 = 0$(CR)	$\alpha^2_{\rm D} = 2/3$(FICD)	$\alpha^2_{\rm P}$(FICP)	$\alpha^2_{\rm M}$(FICM2)
1	8.0000000000	8.0000000000	8.0000000000	8.0000000000	8.0000000000
2	0.1535996812	−1.0514115126	0.4553713752	0	0.1535996812
3	0.0029491079	0.1381982018	0.0259204876	0	0.0029491079
4	0.0000566306	−0.0182784646	0.0014772218	0	0.0000566306
5	0.0000014948	0.0028186197	0.0001154770	0	0.0000014948
6	0.0000212544	−0.0071239995	0.0005580649	0	0.0000212544
7	0.0011059158	0.0518597445	0.0097204167	0	0.0011059158
8	0.0575998805	−0.3942838932	0.1707642790	0	0.0575998805
9	3.0000000000	3.0000000000	3.0000000000	3.0000000000	3.0000000000

Table 2 Ritz-FIC eight-element solutions, $w = 50$

Node	Exact	$\alpha^2 = 0$(CR)	$\alpha^2_{\rm D} = 2/3$(FICD)	$\alpha^2_{\rm P}$(FICP)	$\alpha^2_{\rm M}$(FICM2)
1	8.0000000000	8.0000000000	8.0000000000	8.0000000000	8.0000000000
2	3.3105043651	3.2068850933	3.4002474704	0	3.3105043651
3	1.3801678534	1.2942058990	1.4569382771	0	1.3801678534
4	0.6001402687	0.5439870932	0.6518621127	0	0.6001402687
5	0.3203030826	0.2823794489	0.3560532239	0	0.3203030826
6	0.3074243641	0.2744060401	0.3384109163	0	0.3074243641
7	0.5507702703	0.5129051224	0.5851521371	0	0.5507702703
8	1.2531588628	1.2120974286	1.2890434650	0	1.2531588628
9	3.0000000000	3.0000000000	3.0000000000	3.0000000000	3.0000000000

Results for $w = 50$. This case pertains to moderate absorption to diffusion ratio Eq. 14. The boundary layers are diffuse and the exact solution resembles a second degree parabola. The problem is again discretized using four α choices: $\alpha = 0$, $\alpha^2_{\rm C} = 2/3$, $\alpha^2_{\rm P} = -334/75 = -4.45333$ and $\alpha^2_{\rm M} = 0.345961$. The numerical results are plotted in Fig. 10b and listed in Table 2. Again the solution for $\alpha^2_{\rm M}$ is nodally exact. The solutions for $\alpha = 0$ and $\alpha^2_{\rm C} = 2/3$ bound the exact solution, maintain positivity and display reasonable accuracy. The results for $\alpha_{\rm P}$ are way off as can be expected from the rationale for its construction.

Results for $w = -50$. The solution to this Helmholtz equation goes roughly through one wavelength over the problem domain. Negatives values of $u(x)$ are physically admissible. The problem is discretized with four α choices: $\alpha = 0, \alpha^2_{\rm C} = 2/3, \alpha^2_{\rm P} = 434/75 = 5.78667$ and $\alpha^2_{\rm M} = 0.319898$. Note that $\alpha^2_{\rm M}$ is still positive because $w\chi^2 = -50/64 > -11.4746$,

Table 3 Ritz-FIC eight-element solutions, $w = -50$

Node	Exact	$\alpha^2 = 0$(CR)	$\alpha_{\rm D}^2 = 2/3$(FICD)	$\alpha_{\rm P}^2$(FICP)	$\alpha_{\rm M}^2$(FICM2)
1	8.0000000000	8.0000000000	8.00000000000	8.0000000000	8.0000000000
2	2.1905465154	0.0898920049	3.91683000520	0	2.1905465154
3	−5.2217158313	−7.8823533208	−3.22636343116	0	−5.2217158313
4	−8.8132821321	−10.4059673187	−7.84896043693	0	−8.8132821321
5	−5.9562263170	−5.7365163498	−6.33955710134	0	−5.9562263170
6	1.2589621643	2.8982685015	0.12262521967	0	1.2589621643
7	7.5529760894	9.5296419462	6.48900658781	0	7.5529760894
8	8.3205257619	9.5737052899	7.78585155923	0	8.3205257619
9	3.0000000000	3.0000000000	3.00000000000	3.0000000000	3.0000000000

Table 4 Ritz-FIC eight-element solutions, $w = -1000$

Node	Exact	$\alpha^2 = 0$(CR)	$\alpha_{\rm D}^2 = 2/3$(FICD)	$\alpha_{\rm P}^2$(FICP)	$\alpha_{\rm M}^2$(FICM2)
1	8.0000000000	8.0000000000	8.0000000000	8.0000000000	8.0000000000
2	11.5432046602	−4.5551330632	−0.5903534034	0	11.5432046602
3	−23.8970551078	2.6374205639	0.0435651218	0	−23.8970551078
4	21.3673096334	−1.6039299876	−0.0032213812	0	21.3673096334
5	−5.5295477072	1.1081731645	0.0003261977	0	−5.5295477072
6	−13.7521345403	−0.9839426045	−0.0012230626	0	−13.7521345403
7	24.4686939039	1.1895887560	0.0163380309	0	24.4686939039
8	−19.9456285035	−1.7940565712	−0.2213826078	0	−19.9456285035
9	3.0000000000	3.0000000000	3.0000000000	3.0000000000	3.0000000000

and no imaginary numbers appear. The numerical results are plotted in Fig. 10c and listed in Table 3. Again the solution for $\alpha_{\rm M}^2$ is nodally exact. The results for $\alpha = 0$ and $\alpha_D^2 = 2/3$ bound the exact solution and follow its shape reasonably well. The solution for $\alpha_{\rm P}$ is worthless.

Results for $w = -1{,}000$. The rapidly oscillatory response goes roughly through five cycles over the problem domain, since $\kappa = \sqrt{-w}/(2\pi) \approx 5.03$. The problem is discretized with $\alpha = 0, \alpha_{\rm C}^2 = 2/3, \alpha_{\rm P}^2 = 0.92667$ and $\alpha_{\rm M}^2 = -0.261754$. Here $\alpha_{\rm M}^2$ is negative because $w\chi^2 = -1000/64 = -11.875 < -11.4746$. The numerical results are plotted in Fig. 10d and listed in Table 4. Although $\alpha_{\rm M}^2$ produces a nodally exact solution, an eight-element piecewise linear interpolation over five wavelengths is plainly inadequate to capture intra-element oscillations.

The effect of injecting more elements in the later case is illustrated in Fig. 11. This shows results for three choices of α and 8 through 64 elements. A 64-element mesh places about 10 elements per wavelength, which should be adequate as per well-known empirical rules for approximating sinusoidal waveforms. The beneficial effect of nodal exactness is evident. Results for the other two non-matching choices of α, notably conventional Ritz, display erratic behavior even for fine discretizations.

Replacing the exactly integrated Ritz-FIC model by the reduced-integrated model with the matching $\bar{\alpha}_{\rm M}$ of Eq. 43 yields identical results.

6 Templates

The exact- and reduced-integrated FIC-based elements obtained by setting α^2 and α as per Eq. 36 and Eq. 43, respectively, are nodally exact but lead to completely different finite element matrices. A nodally exact model for the Helmholtz equation, developed by Harari and Hughes [15–17] produces yet another set of matrices. This surprising lack of uniqueness raises the question: how many nodally exact elements can be constructed for the absorption–diffusion-Helmholtz equation with constant coefficients? Templates provide the correct answer: an infinite number.

Templates [7] are parametrized algebraic forms of FEM matrices that include all possible elements once *a priori* constraints, such as symmetry, are enforced. Setting parameters to specific values or functions produces element instances. The set of such values is called the *template signature*.

Assuming symmetry, the element coefficient matrix of the model problem can be placed in the template framework by splitting $\mathbf{S}^e = \mathbf{S}_{\rm K}^e + \mathbf{S}_{\rm M}^e$, with

$$\mathbf{S}_{\rm M}^e = -\frac{w\chi^2}{6L^e}\begin{bmatrix} 2+\beta_1+\beta_2+\beta_3 & 1-\beta_1 \\ 1-\beta_1 & 2+\beta_1+\beta_2-\beta_3 \end{bmatrix},$$
$$\mathbf{S}_{\rm K}^e = \frac{1}{L^e}\begin{bmatrix} 1+\beta_4+\beta_5+\beta_6 & -1-\beta_4 \\ -1-\beta_4 & 1+\beta_4+\beta_5-\beta_6 \end{bmatrix}. \tag{45}$$

Matrix $\mathbf{S}_{\rm K}^e$ (stiffness-like) and $\mathbf{S}_{\rm M}^e$ (mass-like) come from the $(u')^2$ and u^2 terms, respectively, in the variational formulation. In Eq. 45 β_1 through β_6 are dimensionless parameters to be chosen. If all of them vanish the conventional Ritz equations Eq. 21 with piecewise linear shape functions result. Thus the β_i may be interpreted as specifying the deviation from conventional Ritz. (If unsymmetric matrices are permitted, as produced for example by Petrov-Galerkin methods, one more parameter would enter each matrix for a total of eight.)

An obvious simplification is that only the sum $\mathbf{S}_{\rm M}^e + \mathbf{S}_{\rm K}^e$ appears in the FEM equations of the diffusion–absorption and

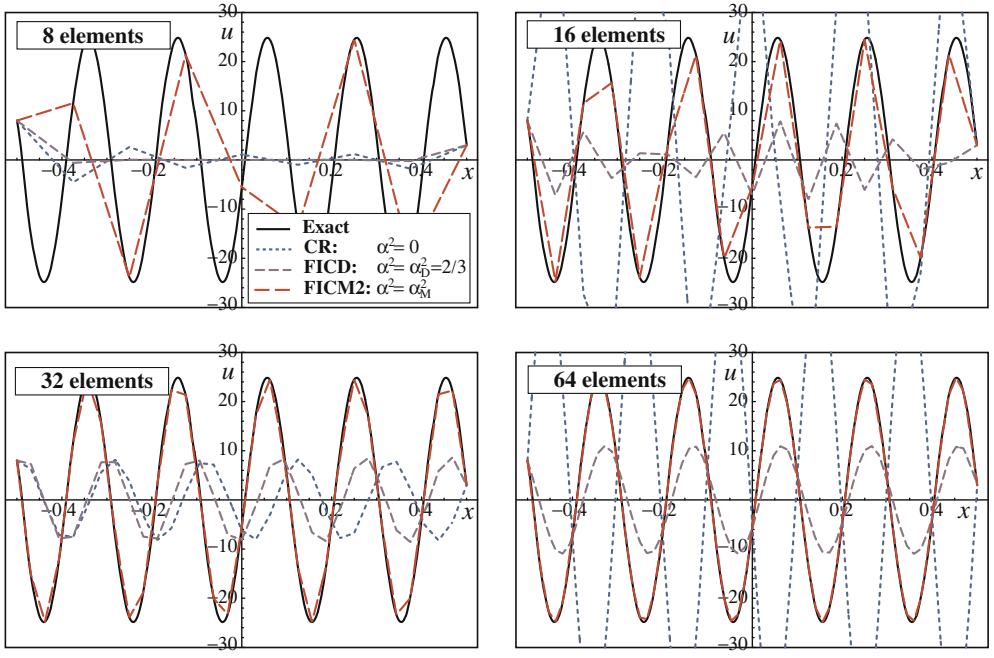

Fig. 11 Ritz-FIC convergence study for rapidly oscillatory case $w = -1000$. Boundary conditions: $u(-\frac{1}{2}) = 8$ and $u(\frac{1}{2}) = 3$. Shown are results for 8, 16, 32 and 64 elements, and three choices of α: α_{M}, α_{D} and conventional Ritz $\alpha = 0$. Results for α_{P} omitted as they are worthless

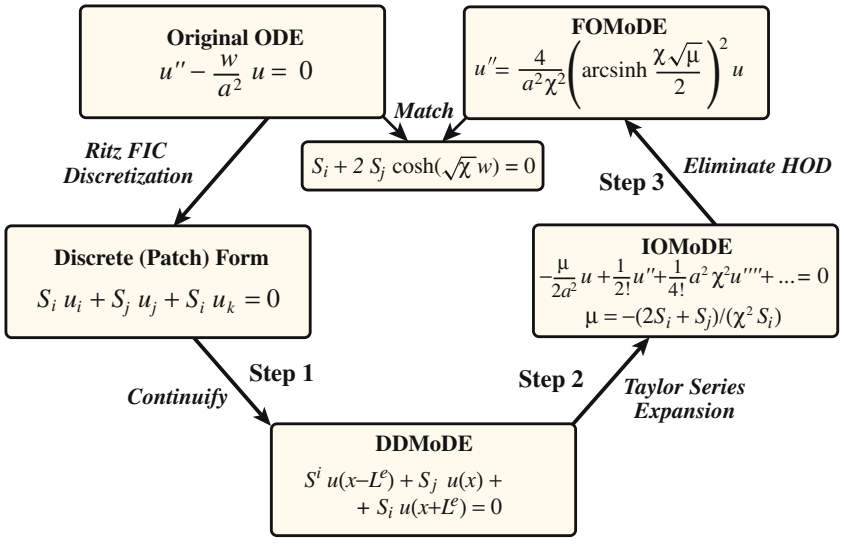

Fig. 12 Nodally exact condition matching for patch equations (Eq. 48) with $S_i = S_k$, which correspond to constant coefficient w and equal length elements $L^e = a\chi$

Helmholtz problems, and three parameters become redundant. Consequently one may either set $\beta_4 = \beta_5 = \beta_6 = 0$ and leave $\mathbf{S}_{\mathrm{K}}$ unchanged, or set $\beta_1 = \beta_2 = \beta_3 = 0$ and leave $\mathbf{S}_{\mathrm{M}}$ unchanged. In the ensuing development the first choice is selected. Thus Eq. 45 reduces to

$$\mathbf{S}^e_{\mathrm{M}} = -\frac{w\chi^2}{6L^e}\begin{bmatrix} 2+\beta_1+\beta_2+\beta_3 & 1-\beta_1 \\ 1-\beta_1 & 2+\beta_1+\beta_2-\beta_3 \end{bmatrix},$$
$$\mathbf{S}^e_{\mathrm{K}} = \frac{1}{L^e}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \qquad (46)$$

If $\beta_2 + 3\beta_1 = 0$ and $\beta_3 = 0$, on taking $\beta_1 = 1 - \beta_0$ and $\beta_2 = 3(\beta_0 - 1)$, β_0 can be extracted as scaling factor of $\mathbf{S}^e_{\mathrm{M}}$ and only one parameter remains:

$$\mathbf{S}^e_{\mathrm{M}} = -\frac{w\chi^2\beta_0}{6L^e}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{S}^e_{\mathrm{K}} = \frac{1}{L^e}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \qquad (47)$$

The general form of the two-element patch equations is

$$S_i\, u_i + S_j\, u_j + S_k\, u_k = 0. \qquad (48)$$

For constant coefficients and equal-length elements, $S_i = S_k$. The modified equation processing steps for this particular

Table 5 Template instances for 1D diffusion–absorption-Helmholtz with constant coefficients

Instance name	Description	Nodally exact?	General performance
CR	Conventional Ritz element Eq. 21	No	Nonphysical oscillations for $w > 0$ if $w\chi^2 > 6$. Erratic behavior for $w < 0$ (Helmholtz)
FICP	Exactly integrated Ritz-FIC element with $\alpha^2 = \alpha_P^2$ of (34), which guarantees positivity if $w > 0$	No	Zero solution at interior nodes, which mimics boundary-layer behavior for huge $w > 0$. Useless for $w < 0$ (Helmholtz)
FICD	Exactly integrated Ritz-FIC element with $\alpha^2 = \alpha_D^2 = 2/3$ of Eq. 35, a setting for high artificial diffusion	No	Correct physical behavior for any $w > 0$, especially very large values. Performs poorly for $w < 0$ (Helmholtz)
FICM1	Reduced integrated Ritz-FIC element with α set as per Eq. 42	CC-ELE	Behavior identical to FICM2 for $w \geq 0$. Higher wavenumber validity range than FICM2 for $w < 0$ (Helmholtz)
FICM2	Exactly integrated Ritz-FIC element with α^2 set as per Eq. 36	CC-ELE	Behavior identical to FICM1 for $w \geq 0$. Imaginary coefficients if $\tilde{\chi}^2 < \zeta^* \approx -11.4746$
HHH	Harari and Hughes [16] nodally exact element for one-dimensional exterior Helmholtz	CC-ELE	Can be used for $w > 0$, where performance is identical to FICM1 and FICM2. For $w < 0$ (Helmholtz) $\mathbf{S}_M$ becomes null at wavenumbers noted in Table 6
NECC	Template set to match FOMoDE for varying length elements	CC-VLE	Performance identical to FICM1, FICM2 and HHH if $w > 0$ and elements are of equal length. Maintains nodal exactness if element lengths vary

CC-ELE: nodally exact for constant coefficients and equal-length elements; *CC-VLE*: nodally exact for constant coefficients and variable length elements

case is diagramed in Fig. 12. The IOMoDE expression displayed there indicates that consistency with the original equation as the mesh is refined requires that $\mu = (2S_i + S_j)/(\chi^2 S_i) \to w$ as $\chi \to 0$. For the template (Eq. 46), $S_i = S_k = 6w(\beta_1 - 1)\chi^2/(6a\chi)$ and $S_j = (6 + 6w(\beta_1 + \beta_2 + 2)\chi^2)/(3a\chi)$. Hence

$$\mu = \left.\frac{2w(3+\beta_2)}{(6+w(\beta_1-1)\chi^2}\right|_{\chi\to 0} = w(1+\tfrac{1}{3}\beta_2) \quad \text{whence} \quad \beta_2|_{\chi\to 0} = 0. \tag{49}$$

Matching for nodal exactness yields $S_i + 2S_j \cosh\zeta = 0$ with $\zeta = \chi\sqrt{w}$. In terms of the parameters of the template Eq. 46, the condition is

$$6 + w\,(2 + \beta_1 + \beta_2)\chi^2 = \left(6 + w\,(\beta_1 - 1)\chi^2\right)\cosh\zeta. \tag{50}$$

Notice that β_3 does not appear in Eq. 50 because of equal-element-length cancellations. The foregoing condition, which may be solved for either β_1 or β_2, shows that there is an *infinite number of nodally exact finite element models* that form a one-parameter family. Four instances of this family are shown in Tables 5 and 6. These are identified by labels FICM1, FICM2, HHH and NECC. The former two are the "matched α" FIC elements obtained in Sects. 5.7 and 5.4, respectively, whereas HHH was obtained by Harari and Hughes [16]. A study of the variable-element-length case, omitted to save space, shows that NECC is the only instance that is nodally exact for such discretizations.

Each nodally exact instance of Tables 6, and 7 has "trouble spots" when applied to the Helmholtz equation $w < 0$, if $\kappa\chi = \chi\sqrt{-w}/(2\pi)$ exceeds 1/2. Those spots are displayed in Fig. 13.

7 Generalization to variable coefficients

This section studies the application of the foregoing discretization methods to the variable-coefficient generalization of Eq. 10 :

$$k\,u'' - s(x)\,u + Q = 0, \tag{51}$$

in which $k \geq 0$ is constant but s is now a linear function of x. As previously, the computational domain is $x \in [x_m, x_p]$ with Dirichlet boundary conditions $u(x_m) = u_m$ and $u(x_p) = u_p$. Of particular interest is when s changes sign over the computational domain. If so $u(x)$ will exhibit boundary-layer exponential decay behavior over the portion where $s > 0$, transitioning to oscillations of varying frequency wherever $s < 0$. This is illustrated by the analytical solution curves in Figs. 4, and 5.

The goal is to investigate whether variational FIC methods can handle simultaneously the diffusion–absorption and Helmholtz type of equations in the same BVP. The restriction to linear variation in x is imposed to have a closed form solution, in terms of Airy functions, available for comparison. Although these functions are rarely useful in classical mechanics they find application in laser optics, quantum mechanics, electromagnetics, and radiative heat transfer.

7.1 The VC model problem

As before we restrict the computational domain to $\pm 1/2a$, set $Q = 0$, and redefine $w(x) = s(x)\,a^2/k$ as a linear function in x explicitly given as $w = w_0 + \psi\,(x/a)$. The model problem is

$$u'' - \frac{w}{a^2}u = 0 \quad \text{with } w = w_0 + \psi\frac{x}{a}, \quad \text{for} x \in [-\tfrac{1}{2}a, \tfrac{1}{2}a], \quad u(-\tfrac{1}{2}a) = u_m, u(\tfrac{1}{2}a) = u_p. \tag{52}$$

Table 6 Template signatures for elements of Table 5

Instance name	Template form	Template signature	Trouble spots
CR	Eq. 46	$\beta_1 = \beta_2 = \beta_3 = 0$	Nonphysical oscillations if $w\chi^2 > 6$
FICP	Eq. 46	$\zeta = \chi\sqrt{w}, \quad \alpha^2 = 2/3 - 4/\zeta^2$ $\beta_1 = 1 - 6/\zeta^2, \quad \beta_2 = 0, \quad \beta_3 = 3\alpha$	$\beta_1 \to \infty$ if $\zeta^2 = w\chi^2 = 6$
FICD	Eq. 46	$\beta_1 = 1, \quad \beta_2 = 0, \quad \beta_3 = \sqrt{6}$	
FICM1	Eq. 46	$\zeta = \chi\sqrt{w}, \quad \tau = \cosh\zeta,$ $\bar{\alpha} = \dfrac{1 - (\tau/\zeta)\sqrt{\zeta^2 - 4 + 8/\tau - 4/\tau^2}}{\tau - 1}$ $\beta_1 = (3\bar{\alpha}^2 - 1)/2, \quad \beta_2 = 3\bar{\alpha}, \quad \beta_3 = 0$	If $w < 0$, entries jump at $\sqrt{-w}\chi = \frac{1}{2}\pi n, \quad n = 1, 2, \ldots$. See Figs. 9 and 13a
FICM2	Eq. 46	$\zeta = \chi\sqrt{w}, \quad \alpha^2 = \frac{2}{3} - \frac{4}{\zeta^2} + \dfrac{1}{\sinh^2(\zeta/2)}$ $\beta_1 = 3\alpha^2/2, \quad \beta_2 = 0, \quad \beta_3 = 3\alpha$	If $w < 0$ and $\chi\sqrt{-w} > 3.3874\ldots$ diagonal entries of $\mathbf{S}^e_{\mathrm{M}}$ become complex. See Figs. 9 and 13b
HHH	Eq. 47	$\zeta = \chi\sqrt{w}, \quad \tau = \cosh\zeta, \beta_0 = \dfrac{6(\tau - 1)}{\zeta^2(\tau + 2)}$	If $w < 0$ and $\chi\sqrt{-w} = 2\pi n, n = 1, 2, \ldots, \beta_0 = 0$, which makes $\mathbf{S}^e_{\mathrm{M}} = \mathbf{0}$. See Fig. 13c
NECC	Eq. 46	$\zeta = \chi\sqrt{w}, \quad \beta_1 = \dfrac{\zeta^2 - 6 + 6\zeta/\sinh\zeta}{\zeta^2},$ $\beta_2 = (6/\zeta)\tanh(\zeta/2) - 3, \quad \beta_3 = 0$	If $w < 0$ and $\chi\sqrt{-w} = \pi n, n = 1, 2, \ldots$, entries of $\mathbf{S}^e_{\mathrm{M}}$ blow up. See Fig. 13d

Table 7 Template instances for 1D diffusion–absorption-Helmholtz with variable coefficients

Instance name	Description	Nodally exact?	General performance
MNECC	Modification of NECC to account approximately for the variable coefficient by matching a truncated MoDE	No	Reasonably accurate for $w > 0$ if variation is not abrupt. Requires fine mesh for Helmhotz
CUBVC	Similar to NELVC with Airy functions approximated by cubic over element	No	Excellent for $w > 0$, even for sharp coefficient variation. Less accurate for Helmholtz but better than MNECC
NELVC	Uses Airy functions to match MoDE exactly	Yes	Nodally exact for any linear coefficient variation, no matter how abrupt. Also exact for variable length elements

Table 8 Template signatures for elements of Table 7

Instance name	Template form	Template signature	Trouble spots
MNECC	Eq. 46	$\beta_1, \beta_2 =$ same as for NECC, $\beta_3 = \dfrac{-3\,\psi\chi(\zeta\coth\zeta - 1)}{2w\zeta^2}$	Same as NECC
CUBVC	Eq. 46	$d = 10w(11760 + 1456w\chi^2 + 28w^2\chi^4 - \chi^6\psi^2)$ $\beta_1 = 7\chi^2(1960w^2 + 36w^3\chi^2 + 14\chi^2\psi^2 - w\chi^4\psi^2)/d$ $\beta_2 = -7(4200w^2\chi^2 + 100w^3\chi^4 + 70\chi^4\psi^2 - 3w\chi^6\psi^2)/d$ $\beta_3 = -3\chi\psi(39200 + 2240w\chi^2 + 28w^2\chi^4 - \chi^6\psi^2)/d$	Same as NECC
NELVC	Eq. 46	$\Upsilon = (\sqrt[3]{\psi^2}), \quad z_m = \frac{(w-(1/2)\chi\psi)}{\Upsilon}, \quad z_p = \frac{(w+(1/2)\chi\psi)}{\Upsilon}$ $A_p = Ai(z_p), \ A_m = Ai(z_m), \ B_p = Bi(z_p), \ B_m = Bi(z_m)$ $A'_p = Ai'(z_p), \ A'_m = Ai'(z_m), \ B'_p = Bi'(z_p), \ B'_m = Bi'(z_m)$ $d_1 = A'_m B_m + A'_p B_p - A_m B'_m - A_p B'_p$ $d_2 = (A'_m - A'_p)(B_m - B_p) - (A_m - A_p)(B'_m - B'_p)$ $d_3 = A'_p B_m - A'_m B_p + A_p B'_m - A_m B'_p, \ d = A_p B_m - A_m B_p$ $\beta_1 = 1 - 6/\zeta^2 + \dfrac{3\,d_1\chi\Upsilon}{d\,\zeta^2}, \ \beta_2 = -3 - \dfrac{3\,d_2\chi\Upsilon}{d\,\zeta^2}, \ \beta_3 = -\dfrac{3\,d_3\chi\Upsilon}{d\,\zeta^2}$	Fails for $\psi = 0$ if not checked

The associated functional is obtained by simply changing w in Eq. 16 to $w(x)$:

$$J[u] = \int_{-a}^{a} \left((u')^2 + \frac{w_0 + \psi x/a}{2a^2} u^2 \right) \mathrm{d}x. \tag{53}$$

where variation is taken over continuous $u(x)$ that satisfy *a priori* the Dirichlet BCs. The solution of Eq. 52 can be expressed in closed form in terms of the Airy functions $Ai(x)$ and $Bi(x)$, as defined for example in [1 , Sect. 10.4], as follows. Assuming that $\psi \neq 0$, compute:

$$\Upsilon = \left(\sqrt[3]{\psi^2} \right), \quad \xi = \frac{2x}{a}, \quad z_\xi = \frac{(w_0 + (1/2)\xi\psi)}{\Upsilon,}$$

$$z_p = \frac{(w_0 + (1/2)\psi)}{\Upsilon}, \quad z_m = \frac{(w_0 - (1/2)\psi)}{\Upsilon,}$$

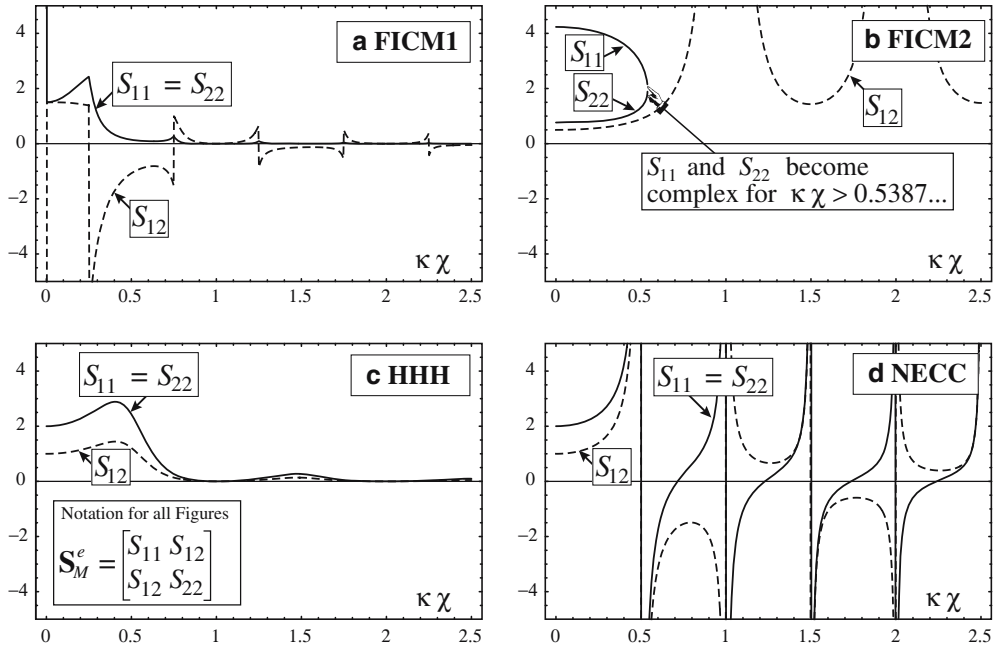

Fig. 13 Trouble spots in coarse, nodally-exact discretizations of the constant-coefficient Helmholtz equation $w < 0$. Graphs plot entries of the 2×2 matrix $\mathbf{S}_{\mathrm{M}}^e$ for instances **a** FICM1, **b** FICM2, **c** HHH and **d** NECC, as functions of the scaled wavenumber $\kappa\chi = \sqrt{-w}\chi/(2\pi)$. The number of elements per wavelength is $N_\lambda^e = 1/(\kappa\chi)$. Difficulties start at roughly $\kappa\chi > 1/2$, or $N_\lambda^e < 2$. Jumps in (**a**) occur at Dirichlet resonances and are harmless

$$
\begin{aligned}
&A_p = Ai(z_p), \quad A_m = Ai(z_m), \quad B_p = Bi(z_p),\\
&B_m = Bi(z_m), \quad A_\xi = Ai(z_\xi), \quad B_\xi = Bi(z_\xi),\\
&d = A_m B_p - A_p B_m,\\
&u(x) = \left[\frac{(B_p A_\xi - A_p B_\xi)}{d}\right] u_m + \left[\frac{(A_m B_\xi - B_m A_\xi)}{d}\right] u_p.
\end{aligned}
\tag{54}
$$

(In symbolic work, it is important to square ψ before taking the cubic root.) If $\psi = 0$ the solution of Eq. 54 fails. In that case $w = w_0$ is constant and the exact solution in terms of exponentials given by Eq. 17, 18 and 19 should be used. As an example, Figs. 4, and 5 show the exact solution of Eq. 9, a BVP that fits Eq. 52 with $w = -3000x/a^2$, $a = 2$, $u(-1) = 25$ and $u(1) = -4$.

The finite-order modified equation for this ODE requires hypergeometric functions to be expressed in closed form, and is omitted for brevity.

7.2 Reusing CC templates

An expedient approach to FEM discretization of Eq. 52 is to reuse the template instances of Tables 5 and 6. If the nodal values of $w(x)$ are w_i and w_j, the average value $w_m = (w_i + w_j)/2$ is used to form the element. See Fig. 14. For example, the reduced-integration Ritz-FIC element FICM1 is again given by Eq. 41 except that $\alpha = \bar{\alpha}_{\mathrm{M}}$ comes from w_m. For the HHH and NECC instances w_m is inserted in the formulas of Table 6.

For the Ritz-FIC element FICM2 a slight refinement is to evaluate $w(x) = w(\xi)$ at the Gauss points $\xi = \pm 1/\sqrt{3}$,

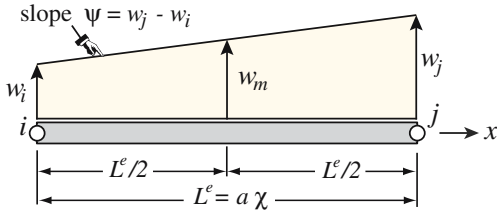

Fig. 14 Variable coefficient variation over element

compute α_{M} there, and use those in the two-point Gauss quadrature. In the numerical experiments, however, the reduced integration FICM1 performed consistently better than FICM2, even after the foregoing refinement was incorporated.

The reuse strategy was found to generally work well over the exponential-behavior portion of the computational domain in which $w > 0$. For oscillatory (Helmholtz) portions in which $w < 0$, convergence was erratic unless a very fine mesh was used. These findings are clearly illustrated in the plots of Figs. 4, and 5. In particular, Fig. 5b shows that 12 elements per shortest wavelength are needed to get satisfactory convergence of FICM1 over the oscillatory region.

7.3 VC-customized templates

In an effort to improve convergence over oscillatory regions, the Ritz-FIC steplength parameter α was allowed to be a

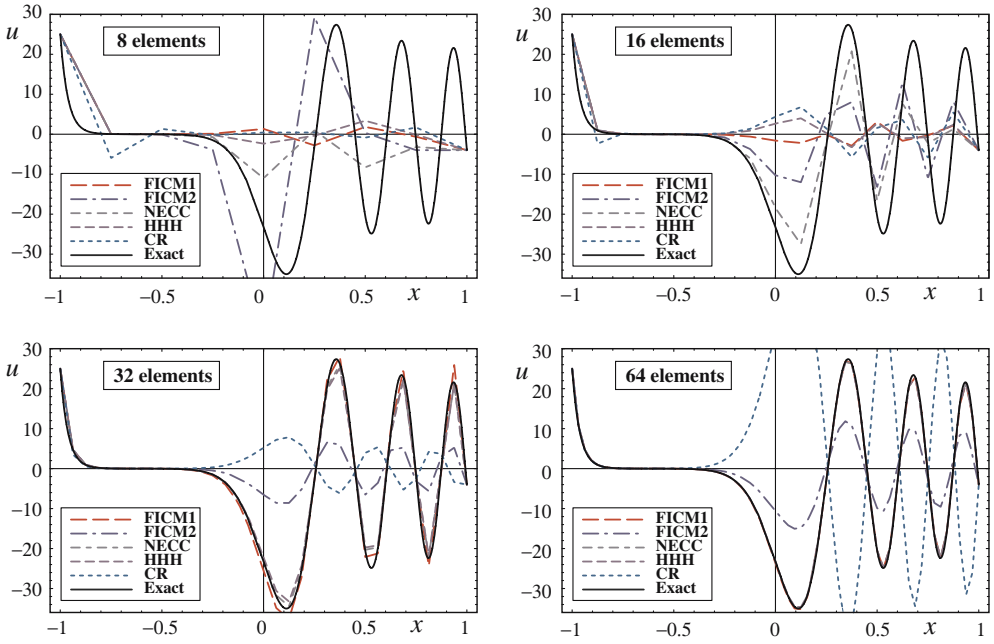

Fig. 15 Convergence study of variable coefficient ODE $u'' = -3000\,x\,u$ with $a = 2$, Dirichlet boundary conditions: $u(-1) = 25$ and $u(1) = -4$. Results shown for 8, 16, 32 and 64 elements, nodally exact templates of Tables 6 and 7 plus CR

function of x: $\alpha = \alpha(x)$, and matching with the modified VC equation attempted. However, no matching was found.

Gradually it was realized was that a single steplength parameter was insufficient, and the more general framework of templates introduced in Sect. 6 was necessary. The developments are lengthy and only the final results are reported. Three useful template instances customized to the VC case are listed in Tables 7 and 8. Model NELVC is nodally exact for linearly varying coefficient. (It is not known whether NELVC is the only nodally exact model for this case). One drawback of NELVC, however, is that its template parameters are expressed in terms of Airy functions and their x-derivatives. These may be difficult to evaluate unless appropriate libraries are available. The other two template instances: MNECC and CUBVC, generally produce better results than those in Tables 5 and 6 for the same mesh, although they are not nodally exact.

7.4 Numerical results for variable coefficients

The test BVP is Eq. 9, i.e. $u'' = -750\,x\,u$, $a = 2$, $u(-1) = 25$ and $u(1) = -4$. Consequently $w = -750\,x\,a^2 = -3000\,x$ varies from 3000 on the left to -3000 at the right, and $\psi = -6000$. Results for meshes with 8 through 64 elements are presented in Figs. 15 and 16.

Figure 15 collects results for the four constant-coefficient nodally exact models of Tables 5, and 6: FICM1, FICM2, HHH and NECC, plus Conventional Ritz (CR). All models perform reasonably well in the exponential region $x < 0$, and capture the boundary layer near $x = -1$ well as the mesh is refined. In the oscillatory region $x > 0$ three models: FICM1, NECC and HHH, start to converge satisfactorily at 32 elements, and agree well with the exact solution at 64 elements. The other two models: CR and FICM2, display erratic behavior throughout; in fact convergence was noticeable only on using 256 elements or more.

Figure 16 collects results for the three VC-customized template instances of Tables 7 and 8: MNECC, CUBVC and NELVC. Again the exponential side is accurately captured even for the coarsest mesh. NELVC is of course nodally exact, and its only deficiency is the intra-element variation. The other two models begin to display nodal convergence at 16 elements, even near $x = 1$, and can hardly be distinguished from NELVC at the plot scale for 32 and 64 elements. MNECC performs surprisingly well considering the simplicity of its template signature, given in Table 8.

8 Conclusions

This article has presented a synthesis of three techniques: FIC, variational Ritz and modified differential equations. The major new contributions are:

1. The FIC approach to functional modification. This permits effective stabilization of the diffusion–absorption problem while staying within the ordinary Ritz framework of finite elements. No separate choice of trial and weight functions is necessary.
2. The use of the modified equation (MoDE) approach to find a value of the stabilization parameter that is nodally exact for all values of the absorption-to-diffusion ratio,

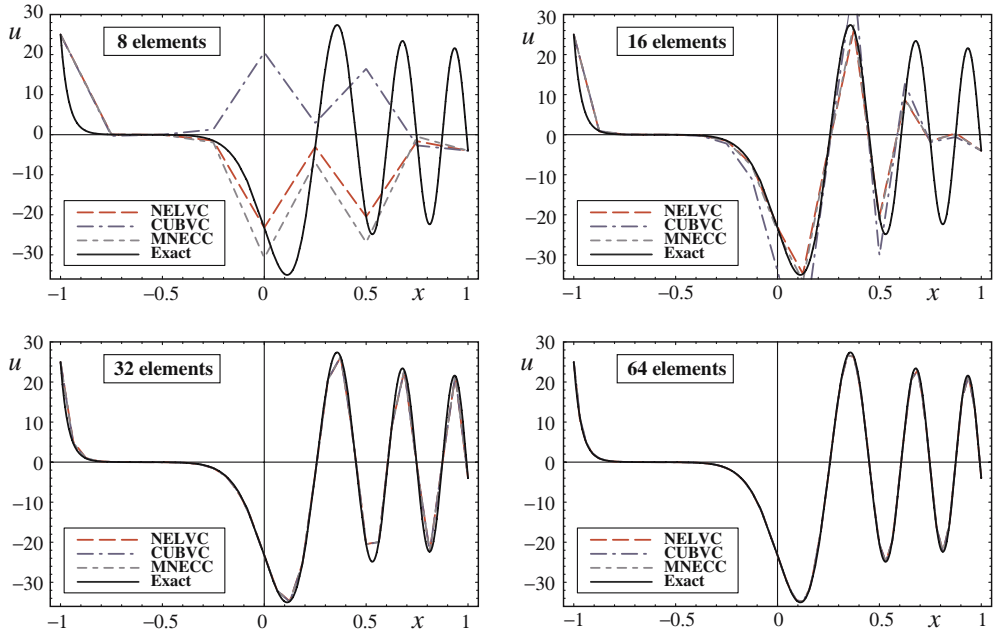

Fig. 16 Convergence study of variable coefficient ODE $u'' = -3000\,x\,u$ with $a = 2$, Dirichlet boundary conditions: $u(-1) = 25$ and $u(1) = -4$. Results shown for 8, 16, 32 and 64 elements, and templates of Tables 8 and 9

including negative values that morph the original ODE into the Helmholtz equation.

One surprise was the discovery that nodally exact discretizations for the constant coefficient case are not unique, which explains variants in the published literature. With the introduction of templates as described in Sect. 6, all such instances can be characterized once and for all. Templates also allowed the variable coefficient case to be tamed, although uniqueness remains an open question.

The chief attraction of the modified equation approach is that availability of exact solutions of the source ODE is not required to construct accurate discretizations. This feature is important for application of the method in two and three dimensions. For the one-dimensional problem discussed here, nodally exact discretization can be also obtained by patch matching as illustrated in Sect. 5.6.

The logical extension of the present combination of methods is the study of two- and three-dimensional space discretizations by considering regular finite element patches. Since exact solutions for such problems are rarely available, the modified equation method appears to be a promising choice for improving nodal solutions over fixed meshes. The Ritz ingredient, however, may have to be dropped in problems, such as advection, that are not easily formulated in a variational framework

Appendix A Processing modified equations

This Appendix presents two mathematical procedures that find application in the modified equation method. Technically the most difficult operation in the process of Fig. 3 is passing from the infinite order modified equation (IOMoDE) to a finite order one (FOMoDE). There is no universal method for doing this reduction because the process involves identification of series. Case by case is the rule. Nonetheless there are some differential equations of mathematical physics that naturally lead to Toeplitz matrix forms. If this happens, an array of powerful techniques is available. This is illustrated by an example that conveys the flavor of the method as well as finding applications in the matching procedures for nodal exactness of Sects. 5 and 6.

A.1 Reduction by series identification

Consider the homogeneous, even-derivative, infinite-order ODE in the dependent variable $u(x)$:

$$-\frac{\mu}{2a^2}u(x) + \frac{1}{2!}u''(x) + \frac{a^2\chi^2}{4!}u''''(x) + \frac{a^4\chi^4}{6!}u''''''(x) + \cdots = 0, \quad a > 0, \quad \mu \neq 0, \quad 0 < \chi \leq 1. \tag{55}$$

Here μ and χ are dimensionless real parameters whereas a, which is a characteristic problem dimension, has dimension of length. The reduction to finite order can be obtained by a variant of Warming and Hyett's [46] derivative elimination procedure, Differentiate (55) $2(n-1)$ times ($n = 1, 2, \ldots$) with respect to x while discarding all odd derivatives. Truncate to the same level in χ, and set up a linear system in the even derivatives u'', u'''', ... The configuration of the elimination system is illustrated for $n = 4$:

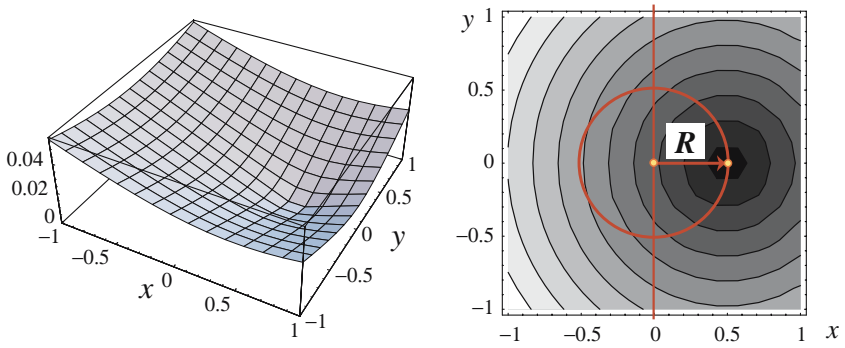

Fig. 17 Plot of the modulus $\sqrt{g_r^2 + g_i^2}$ of generating function (64) for $\{\mu = 1/2, a = 1, \chi = 1/4\}$, showing the only zero at $x \approx 0.4998$, $y = 0$ and the convergence radius R

$$\begin{bmatrix} \frac{1}{2!} & \frac{a^2\chi^2}{4!} & \frac{a^4\chi^4}{6!} & \frac{a^6\chi^6}{8!} \\ -\frac{1}{2}\mu a^{-2} & \frac{1}{2!} & \frac{a^2\chi^2}{4!} & \frac{a^4\chi^4}{6!} \\ 0 & -\frac{1}{2}\mu a^{-2} & \frac{1}{2!} & \frac{a^2\chi^2}{4!} \\ 0 & 0 & -\frac{1}{2}\mu a^{-2} & \frac{1}{2!} \end{bmatrix} \begin{bmatrix} u'' \\ u'''' \\ u'''''' \\ u'''''''' \end{bmatrix} = \begin{bmatrix} \frac{1}{2}a^{-2}\mu\, u \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{56}$$

The coefficient matrix of this system is Toeplitz and Hessemberg but not Hermitian. This can be solved for u'' to yield a truncated FOMoDE. Solving Eq. 56 and expanding in Taylor series gives

$$u'' = \frac{56\mu(360 + 60\lambda + \lambda^2)\, u}{20160 + 5040\lambda + 252\lambda^2 + \lambda^3} = \mu\left(1 - \frac{1}{12}\lambda + \frac{1}{90}\lambda^2 + \cdots\right) u. \tag{57}$$

where $\lambda = a^2\chi^2\mu$. Increasing n, the coefficients of the power series in λ are found to be generated by the recursion $c_1 = 1$, $c_{k+1} = -\frac{1}{2}k^2 c_k/[(k+1)(2k+1)], k \geq 1$, which produces the sequence $\{1, -1/12, 1/90, -1/560, 1/3150, -1/16632, \ldots\}$. The generating function [46] can be found by *Mathematica's* package RSolve by entering «DiscreteMath'RSolve'; g=GeneratingFunction [a[k+1]==–k*k/(2*(k+1)*(2*k+1))* a[k], a[1]==1, a[k],k,λ]; Print[g]. The answer may be verified by Print [Series[g,{ λ,0,8 }]. The result is

$$\frac{4}{\lambda}\left(\operatorname{arcsinh}\frac{\sqrt{\lambda}}{2}\right)^2 = 1 - \frac{\lambda}{12} + \frac{\lambda^2}{90} - \frac{\lambda^3}{560} + \frac{\lambda^4}{3150} - \frac{\lambda^5}{16632} + \frac{\lambda^6}{84084} - \frac{\lambda^7}{411840} + \cdots \tag{58}$$

This yields the second-order FOMoDE

$$u'' = \frac{4}{a^2\chi^2}\left(\operatorname{arcsinh}\frac{\sqrt{\lambda}}{2}\right)^2 u = \frac{4}{a^2\chi^2}\left(\operatorname{arcsinh}\frac{\chi\sqrt{\mu}}{2}\right)^2 u. \tag{59}$$

Suppose that the original ODE is that of the model BVP Eq. 15 : $u'' = (w/a^2)\, u$, with constant w. For nodal exactness, $w = (4/\chi^2)\big(\operatorname{arcsinh}(\frac{1}{2}\chi\sqrt{\mu})\big)^2$. If μ is the free parameter, solving for it gives

$$\mu = \frac{4}{\chi^2}\left(\sinh\frac{\chi\sqrt{w}}{2}\right)^2 = \frac{2\big(\cosh(\chi\sqrt{w}) - 1\big)}{\chi^2}. \tag{60}$$

This is used in Sect. 5.4 to produce Eq. 36. In the foregoing analysis no term of the Eqs. 55 or 59 is assumed to be small. The procedure for handling a forcing term $f(x)$ follows essentially the same technique.

Occasionally it is useful to recover higher derivatives in terms of u. This is done by repeated differentiation of the FOMoDE. For this example, if $u'' = C\, u$, then $u'''' = C\, u'' = C^2\, u$, and likewise for higher derivatives.

A.2 Reduction by a theorem of Muir

The construction of Eq. 59 has a heuristic flavor: it relies on recognizing a series. A more direct derivation that however requires more advanced mathematical tools, is presented here. The method relies on the following theorem on determinant recurrences [21 , p. 704]. Suppose that the smooth generating function $g(z)$, where z is complex, has the formal Taylor series $a_0 + a_1 z + a_2 z^2 + \cdots$ at $z = 0$. The reciprocal $1/g(z)$ has the formal expansion

$$\frac{1}{g(z)} = \frac{1}{a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots} = A_0 - A_1 z + A_2 z^2 - A_3 z^3 + \cdots, \quad a_0 \neq 0, \tag{61}$$

Then the Toeplitz determinants formed with the a_i coefficients satisfy

$$A_1 = a_0^{-1}\,|a_1|\,, \quad A_2 = a_0^{-2}\begin{vmatrix} a_1 & a_2 \\ a_0 & a_1 \end{vmatrix}, \quad A_3 = a_0^{-3}\begin{vmatrix} a_1 & a_2 & a_3 \\ a_0 & a_1 & a_2 \\ 0 & a_0 & a_1 \end{vmatrix}, \quad A_4 = a_0^{-4}\begin{vmatrix} a_1 & a_2 & a_3 & a_4 \\ a_0 & a_1 & a_2 & a_3 \\ 0 & a_0 & a_1 & a_2 \\ 0 & 0 & a_0 & a_1 \end{vmatrix}, \ldots \tag{62}$$

with $A_0 = 1/a_0$. Now the determinants that appear in the truncated Toeplitz $n \times n$ matrices in the derivative elimination process, exemplified by Eq. 56 for $n = 4$, have the form

$$A_1 = \left|\frac{1}{2!}\right|,$$

$$A_2 = \begin{vmatrix} \frac{1}{2!} & \frac{a^2\chi^2}{4!} \\ -\frac{1}{2}\mu a^{-2} & \frac{1}{2!} \end{vmatrix},$$

$$A_3 = \begin{vmatrix} \frac{1}{2!} & \frac{a^2\chi^2}{4!} & \frac{a^4\chi^4}{6!} \\ -\frac{1}{2}\mu a^{-2} & \frac{1}{2!} & \frac{a^2\chi^2}{4!} \\ 0 & -\frac{1}{2}\mu a^{-2} & \frac{1}{2!} \end{vmatrix}, \ldots \tag{63}$$

The n^{th} truncated-Toeplitz approximation to the FOMoDE ($n > 1$) is $u'' = C_n u$, with $C_n = A_{n-1}/A_n$. If the series Eq. 61 has radius of convergence R, then $C_n \to 1/R$ as $n \to \infty$. From Eq. 61 through Eq. 63 one obtains by inspection

$$a^2\chi^2\, g(z) = \cosh(a\chi\sqrt{z}) - (1 + \tfrac{1}{2}\mu\chi^2). \tag{64}$$

The radius of convergence of $1/g(z)$ is the distance from $z = 0$ to its closest pole, or what is the same, to the zero of smallest modulus of $g(z)$. As pictured in Fig. 17, the function $g(z)$ with $z = x + iy$, has a single zero $x = R$ on the real line $y = 0$. This is obtained by solving $g(R) = 0$ or, equivalently, $\cosh(a\chi\sqrt{R}) = 1 + 1/2\mu\chi^2$, whence $a\chi\sqrt{R} = \operatorname{arccosh}(1 + 1/2\mu\chi^2)$. See Fig. 17 for a geometric interpretation. For $\mu > 0$ this is equivalent to $R = 4a^{-2}\chi^{-2}\big(\operatorname{arcsinh}(1/2\chi\sqrt{\mu})\big)^2$, which leads to the same solution: $u'' = Ru = 4a^{-2}\chi^{-2}\big(\operatorname{arcsinh}(1/2\chi\sqrt{\mu})\big)^2$ found in Eq. 59. This method bypasses determinant expansions and series identification, but is restricted to Toeplitz-Hessemberg matrices.

Acknowledgement The work of the first author has been partly supported through a fellowship awarded by the Spanish Ministerio of Educación y Cultura to visit CIMNE during May–June 2002, and partly by the National Science Foundation under grant High-Fidelity Simulations for Heterogeneous Civil and Mechanical Systems, CMS-0219422. Thanks are due Professor Hughes for calling attention to the nodally exact solutions in [16,17].

References

1. Abramowitz M, Stegun IA (eds) (1972) Handbook of mathematical functions with formulas, graphs, and mathematical tables, 9th edn. Dover, NY
2. Ahmed MO, Corless RM (1997) The method of modified equations in Maple. In: Electronic proceedings of 3rd international IMACS conference on applications of computer algebra, Maui, Hawaii. PS version accessible at www.math.unm.edu/ACA/1997/Proceedings/odes/ Ahmed_paper.ps
3. Brezzi F, Hughes TJR, Marini LD, Russo A, Süli E (1999) A priori error analysis of residual-free bubbles for advection-diffusion problems. SIAM J Numer Anal 36:1933–1948
4. Brezzi F, Russo A (2000) Stabilization techniques for the finite element method. In: Spigler R (ed) Applied and industrial mathematics venice-2, 1998. Kluwer, Dordrecht, pp 47–58
5. Farhat C, Harari I, Franca L (2001) The discontinuous enrichment method. Comput Methods Appl Mech Eng 190:6455–6479
6. Farhat C, Harari I, Hetmaniuk U (2003) The discontinuous enrichment method for multiscale analysis. Comput Methods Appl Mech Eng 192:3195–3209
7. Felippa CA (2004) A template tutorial. In: Mathisen KM, Kvamsdal T, Okstad KM (eds) Computational mechanics: theory and practice, CIMNE, Barcelona, pp 29–68
8. Finlayson BM (1972) The methods of weighted residuals and variational principles. Academic Press, NY
9. Franca L, Farhat C, Lesoinne M, Russo A (1998) Unusual stabilized finite element methods and residual-free-bubbles. Int J Numer Methods Fluids 27:159–168
10. Franca L, Farhat C, Macedo AP, Lesoinne M (1997) Residual-free bubbles for the Helmholtz equation. Int J Numer Methods Eng 40:4003–4009
11. Griffiths D, Sanz-Serna J (1986) On the scope of the method of modified equations. SIAM J Sci Stat Comput 7:994–1008
12. Kolesnikov A, Baker AJ (1999) Efficient implementation of high order methods for the advection-diffusion equation. Proceedings of 3^{rd} ASME/JSME joint fluids engineering conference. San Francisco, CA
13. Hairer E (1994) Backward analysis of numerical integrators and symplectic methods. Ann Numer Math 1:107–132
14. Hairer E, Lubich C, Wanner G (2002) Geometric numerical integration: structure preserving algorithms for ordinary differential equations. Springer, Berlin Heidelberg New York
15. Harari I, Hughes TJR (1990) Design and analysis of finite element methods for the Helmholtz equation in exterior domains. Appl Mech Rev 43:S366–S373
16. Harari I, Hughes TJR (1991) Finite element methods for the Helmholtz equation in an exterior domain: model problems. Comput Methods Appl Mech Eng 87:59–96
17. Harari I, Hughes TJR (1992) Galerkin/least-square finite element methods for the reduced wave equation with no reflecting boundary conditions in unbounded domains. Comput Methods Appl Mech Eng 98:411–454
18. Hirt CW (1968) Heuristic stability theory for finite difference equations. J Comp Phys 2:339–342
19. Hughes TJR (1995) Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. Comput Methods Appl Mech Eng 127:387–401
20. Kloeden PE, Palmer KJ (eds) (1994) Chaotic Numerics. Amer Math Soc, Providence, RI
21. Muir T (1960) Theory of determinants. Dover, NY
22. Lomax H, Kutler P, Fuller FB (1970) The numerical solution of partial differential equations governing convection. AGARDograph 146-70, NATO Advisory Group for Aerospace Research, Brussels
23. Oñate E, García J, Idelsohn SR, (1997) Computation of the stabilization parameter for the finite element solution of advective-diffusive problems. Int J Numer Methods Fluids 25:1385–1407
24. Oñate E, García J, Idelsohn SR (1998) Computation of the stabilization parameter for the finite element solution of advective-diffusive problems. In: Ladevèze P, Oden JT (eds) New advances in adaptive computer methods in mechanics. Elsevier, Amsterdam, NL
25. Oñate E (1998) Derivation of the stabilization equations for advective-diffusive fluid transport and fluid flow problems. Comput Methods Appl Mech Eng 151:233–267
26. Oñate Manzan M (1999) A general procedure for deriving stabilized space-time finite element methods for advective-diffusive problems. Int J Numer Methods Eng 31:203–207
27. Oñate E, Manzan M (2000) Stabilization techniques for finite element analysis of convection-diffusion fluid problems. In: Comini G, Sunden B (eds) Computerised analysis of heat transfer. WIT Press, Southampton, UK
28. Oñate E (2000) A stabilized finite element method for incompressible viscous flows using a finite increment calculus formulation. Comput Methods Appl Mech Eng 182:355–370
29. Oñate E, García J (2001) A finite element method for fluid-structure interaction with surface waves using a finite calculus formulation. Comput Methods Appl Mech Eng 191:635–660

30. Oñate E (2003) Multiscale computational analysis in mechanics using finite calculus: an introduction. Comput Methods Appl Mech Eng 192:3043–3059
31. Oñate E, Taylor RL, Zienkiewicz OC, Rojek J (2003) A residual correction method based on finite calculus. Eng Comput 20:629–658
32. Oñate E, Taylor RL, Zienkiewicz OC, Rojek J (2004) Finite calculus formulation for analysis of incompressible solids using linear triangles and tetrahedra. Int J Numer Methods Eng 59:1473–1500
33. Oñate E (2004) Possibilities of finite calculus in computational mechanics. Int J Numer Methods Eng 60:255–281
34. Oñate E, García J, Idelsohn SR (2004) Ship Hydrodynamics. In: Hughes TJR, de Borst R, Stein E (eds) Encyclopedia of computational mechanics, vol 2. Wiley, Chichester, pp 579–610
35. Oñate E, Miquel J, Hauke G (2005) A stabilized finite element method for the one-dimensional advection diffusion-absorption equation using finite calculus. Comput. Meth. Appl. Mech. Eng. accepted
36. Oñate E, Felippa CA (2005) Variational formulation of the finite calculus equations in solid mechanics, CIMNE Report (in preparation)
37. Park KC, Flaggs DL (1984) A Fourier analysis of spurious modes and element locking in the finite element method. Comput Methods Appl Mech Eng 42:37–46
38. Park KC, Flaggs DL (1984) An operational procedure for the symbolic analysis of the finite element method. Comput Methods Appl Mech Eng 46:65–81
39. Richtmyer RL, Morton, KW (1967) Difference methods for initial value problems, 2nd edn. Interscience, Wiley, NY
40. Roache PJ (1970) Computational fluid mechanics. Hermosa, Albuquerque, NM
41. Stuart AM, Humphries AR (1996) Dynamic systems and numerical analysis. Cambridge University Press, Cambridge, UK
42. Tong P (1969) Exact solution of certain problems by the finite element method. J AIAA 7:179–180
43. Vujanovic BD, Jones SE (1989) Variational methods in nonconservative phenomena. Academic Press, NY
44. Waltz JE, Fulton RE, Cyrus NJ (1968) Accuracy and convergence of finite element approximations. In: Proceedings of 2nd conference on matrix methods in structural mechanics. WPAFB, Ohio. AFFDL TR 68-150, pp 995–1028
45. Warming RF, Hyett BJ (1974) The modified equation approach to the stability and accuracy analysis of finite difference methods. J Comput Phys 14:159–179
46. Wilf HS (1991) Generating functionology. Academic Press, NY
47. Wilkinson JH (1961) Error analysis of direct methods of matrix inversion. J ACM 8:281–330
48. Wilkinson JH (1963) Rounding errors in algebraic processes. Prentice-Hall, Englewood Cliffs, NJ
49. Wilkinson JH (1965) The algebraic eigenvalue problem. Oxford University Press, Oxford, UK
50. Zienkiewicz OC, Taylor RE (1988) The finite element method, vol I, 4th edn. McGraw-Hill, London, UK