A Class of discontinuous Petrov-Galerkin methods. Part IV: The optimal test norm and time-harmonic wave propagation in 1D

J. Zitelli[a], I. Muga[a,b], L. Demkowicz[a], J. Gopalakrishnan[c], D. Pardo[d], V. M. Calo[e]

[a] *Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78712, USA*

[b] *Instituto de Matemáticas, Pontificia Universidad Católica de Valparaíso, Chile*

[c] *Department of Mathematics, University of Florida, Gainesville, FL 32611–8105, USA*

[d] *Department of Applied Mathematics, Statistics, and Operational Research, University of the Basque Country (UPV/EHU) and Ikerbasque (Basque Foundation for Sciences), Bilbao, Spain.*

[e] *Applied Mathematics and Computational Science, Earth and Science Engineering, King Abdullah University of Science and Technology (KAUST), Saudi Arabia*

Abstract

The phase error, or the pollution effect in the finite element solution of wave propagation problems, is a well known phenomenon that must be confronted when solving problems in the high-frequency range. This paper presents a new method with *no phase errors* for one-dimensional (1D) time-harmonic wave propagation problems using new ideas that hold promise for the multidimensional case. The method is constructed within the framework of the Discontinuous Petrov-Galerkin (DPG) method with optimal test functions. We have previously shown that such methods select solutions that are the best possible approximations in an energy norm dual to any selected test space norm. In this paper, we advance by asking what is the *optimal test space norm* that achieves error reduction in a given energy norm. This is answered in the specific case of the Helmholtz equation with L^2-norm as the energy norm. We obtain uniform stability with respect to the wave number. We illustrate the method with a number of 1D numerical experiments, using discontinuous piecewise polynomial *hp* spaces for the trial space and its corresponding optimal test functions computed approximately and locally. A 1D theoretical stability analysis is also developed.

Key words: time harmonic, wave propagation, Helmholtz, DPG, Discontinuous Petrov Galerkin, robustness, phase error, dispersion, high frequency
2000 MSC: 65N30, 35L15

1. Introduction

The aim of this paper is to introduce a new methodology to design schemes for wave-propagation problems. It is a continuation of our research on Discontinuous Petrov-Galerkin (DPG) methods [10, 11, 12]. Our previous papers applied the DPG methodology to get new methods for convective and diffusive phenomena. In this paper, we apply it to wave propagation after developing additionally needed theoretical tools.

Email addresses: jzitelli@ices.utexas.edu (J. Zitelli), muga@ices.utexas.edu (I. Muga), leszek@ices.utexas.edu (L. Demkowicz), jayg@ufl.edu (J. Gopalakrishnan), dzubiaur@gmail.com (D. Pardo), victor.calo@kaust.edu.sa (V. M. Calo)

The numerical solution of wave propagation problems at high frequencies has been recognized as an outstanding challenge in numerical analysis. In general, numerical methods for wave propagation are subject to the effect of *pollution*: increasing the frequency, while maintaining the approximation quality of the numerical discretization, results in a divergence of the computed result from the best approximation the discretization is capable of. In the context of finite element methods, the pollution error may be characterized as follows [24]: Given that the exact solution u lies in a space U normed by $\|\cdot\|_U$, and the discrete solution u_h in an approximation subspace $U_h \subset U$, one observes that

$$\frac{\|u - u_h\|_U}{\|u\|_U} \leq C(k) \inf_{w_h \in U_h} \frac{\|u - w_h\|_U}{\|u\|_U},$$

where

$$C(k) = C_1 + C_2 k^\beta (kh)^\gamma,$$

with k being the wavenumber, and h being the element size. The infimum measures the best approximation error. This is typically small when kh is small, i.e., when enough elements per wavelength are used. Additional kh dependence may arise through γ. However, more troublesome is the k-dependence in $C(k)$ measured by β. It reflects the growing instability of the problem even before discretization, i.e. the inf-sup constant decreases as k increases. Generally, the exponent β is found to be one [24, 26] – in other words, the "pollution" term in the error increases linearly with frequency. For many model problems, the pollution is manifested as a phase error which accumulates over the domain, and the concepts of pollution, phase error, and discrete wavenumbers are therefore all closely related. The growth of the pollution error, combined with the already difficult problem of approximating the highly oscillatory solutions of wave problems, can render the numerical solution extremely expensive for high wavenumbers.

The main result of our application of the DPG methodology to one-dimensional wave propagation is a Petrov-Galerkin method which is free of pollution, i.e. $\beta = 0$. Additionally, our method also has $\gamma = 0$. A number of previous methods have achieved zero β in 1D, while reducing the severity of the pollution error in higher dimensions. One can find surveys of such methods in, e.g., [28, 20]. Broadly, they may be classified as follows: Galerkin/Least-Squares based methods [21, 29], which achieve improved stability by adding least squares residual terms to the standard Galerkin sesquilinear form; methods utilizing specialized, under-integrating quadrature rules [1] which reduce the phase error, as indicated by dispersion analysis of an interior stencil; and methods incorporating exact solutions of the Helmholtz equation (in particular, plane waves) within the trial space basis [3, 15, 16, 17, 22].

Petrov-Galerkin (PG) formulations also appear frequently in the construction of stabilized methods (see, e.g. [13, 14, 23]). Common to such methods is the introduction of local problems which are solved to provide a trial/test space pair which provides enhanced stability. A few of these methods have attempted to address in particular the Helmholtz equation.

In the nearly optimal Petrov-Galerkin method (NOPG) of Barbone and Harari [5], the authors construct a method with the goal of achieving the best approximation in the H^1 seminorm in a given trial space. They show that the corresponding minimization problem leads to a Petrov-Galerkin formulation with optimal test functions with global support. Then, by considering only local test functions constructed by adding bubbles to the standard basis functions, they arrive at a more practical formulation which approximates the H^1-optimal result. For rectangular/hexahedral elements, the bubble functions may be determined analytically; more generally, the bubbles may be approximated numerically through local Galerkin problems. In certain cases, the method is equivalent to that of residual-free bubbles [18].

The quasi-optimal Petrov-Galerkin (QOPG) method of Loula and Fernandes [25] considers test functions constructed from a linear combination of standard bilinear Lagrangian basis functions and additional bubbles which are products of the same basis functions. The test functions are determined by solving locally a least-squares problem attempting to minimize a residual corresponding to the Lagrange interpolant of plane waves of all directions. For a uniform mesh, the phase error determined by analysis of an interior stencil is of the same order as that of the Quasi-Stabilized FEM (QSFEM) of Babuska et al. [2], i.e.,

$$\frac{|k - k_h|}{k} \leq 1.5 \left(\frac{(kh)^6}{774144} \right), \tag{1.1}$$

where k_h is a "discrete" wave number.

In general, both (NOPG and QOPG) methods require simple preprocessing techniques which can be implemented in existing FEM codes with little extra computational cost. However, both methods fit within the class of generalized finite elements methods (GFEM) analyzed in [4] when restricted to structured meshes. Therefore, we know that in 2D they perform (in the best case) with the same order of phase error as the optimal result [2], i.e., the expression (1.1).

The method we present for Helmholtz problems is very similar in spirit to these other approaches, i.e., it attempts to achieve optimal results in some sense by local computation of corresponding optimal test functions. The use of the DPG setting is where we depart. We have developed such formulations together with the concept of optimal test functions in [10, 11, 12] for convective problems (DPG variational formulations were also considered in [6], but their objective was not to find the best possible test space). Rather than starting from a traditional H^1 variational formulation in terms of pressure, the DPG setting introduces a mixed formulation for both pressure and velocity, which are now in L^2, as well as additional fluxes. We then aim for test functions that yield the best trial approximations in the L^2 norm for both pressure and velocity. The mixed formulation and the discontinuity of the functional spaces is needed to derive an easy, practical, and inexpensive way to compute the optimal test space. Versus the other PG approaches (e.g. [5] or [25]), the method may be difficult to implement within existing classical FEM codes, but fits perfectly within the framework of hybrid methods like the original DPG method developed in [6]. The essential difference is in the computation of optimal test functions, an operation performed purely on the element level using a simple preprocessing routine. Additionally, for a low price, our method also obtains local error indicators for an *hp*-adaptive algorithm (see [12]).

The crucial property of the DPG methodology is that it guarantees the *best approximation property* in the so-called energy (dual or residual) norm [11]. This norm is problem-dependent – it is implied by the operator governing the problem *and the choice of the test space norm*. In our study on convection-dominated diffusion problems presented in [11, 12], our choices of the norms for the test spaces consisted of standard Sobolev norms modified by additional weights to ensure robustness of the resulting method with respect to diffusion coefficient. We then introduced mesh-dependent factors to counteract round-off errors. In this work, we take a different approach and introduce a problem-dependent *optimal test norm*, constructed specifically to obtain a desired energy norm, e.g., the L^2 norm for the problems we consider. Employing the exact optimal test functions corresponding to this test norm therefore yields a method which achieves the best approximation error in the L^2 norm. However, these test functions have global support and we find ourselves in a situation much like that pointed out in [19], i.e., although evaluation of the left hand side matrix is straightforward (by construction, it corresponds to the L^2 inner product), the work has been moved to determining the global optimal test functions which are needed to define the load vector. We therefore adapt our approach to design an equivalent test

space norm which possesses local optimal test functions. This new test norm is equivalent to the optimal one *uniformly in wavenumber* k. Thus we are able to prove quasioptimality of our method with a wavenumber independent constant.

This paper is structured as follows. First we introduce an abstract framework for the method (Section 2), related to the notions of *optimal test norm*, equivalent norms and the DPG method implementability. Next we apply the framework to two time-harmonic wave propagation problems, starting with a simple time-harmonic transport problem (Section 3) and continuing with the Helmholtz equation in a first-order setting (Section 4). Both problems are illustrated with extensive numerical examples. For the Helmholtz equation case, we additionally combine our method with a PML truncation. Conclusions are presented in Section 5 and the appendix section A collects the proofs of technical lemmas used in the text.

2. Petrov-Galerkin Method with Optimal Test Norm

The new methodology for constructing schemes is introduced in this section. Here we recall the DPG framework [10, 11] and the concept of optimal test functions introduced in [11]. This is presented together with the new concept of optimal test norm. In later sections we apply these abstract results to specific wave propagation examples.

2.1. Abstract setting

Consider an arbitrary abstract variational problem,

$$\begin{cases} \text{Find } u \in U \text{ such that :} \\ b(u,v) = l(v), \quad \forall v \in V. \end{cases} \tag{2.2}$$

Here U, V are two reflexive Banach spaces over $\mathbb{C}$ (the complex field), $b(u,v)$ is a continuous sesquilinear form on $U \times V$ and $l(v)$ is a continuous conjugate linear form on V representing the load. (This terminology is standard – see [30]. Conjugate linear forms have also been called antilinear forms [27]).

We denote by U' the space of continuous linear functionals on U and by V^* the space of continuous conjugate linear functionals on V. The sesquilinear form b generates two continuous operators, B and B', defined by

$$\begin{array}{lll} B : U \to V^* & \text{such that} & Bu(v) = b(u,v), \quad \forall u \in U,\ \forall v \in V, \\ B' : V \to U' & \text{such that} & B'v(u) = b(u,v), \quad \forall u \in U,\ \forall v \in V. \end{array}$$

The conjugate operator of a linear operator $L : E \to F$ is $L^* : F' \to E'$, defined by $L^* f(e) = f(Le)$ for all $f \in F'$ and $e \in E$. Since V is reflexive, there is an invertible (conjugate linear) isometry $I_V : V \to (V^*)'$ such that $I_V v(v^*) = v^*(v)$ for all $v \in V$ and $v^* \in V^*$. It is easy to check that

$$B^* \circ I_V = B'. \tag{2.3}$$

We assume now that the operator B is invertible with continuous inverse $B^{-1} : V^* \to U$. We also assume that the operator B' is injective, which implies that it also has a continuous inverse (see e.g. [9, 27]). Then the problem (2.2) is clearly well-posed. Moreover, since $(B^*)^{-1} = (B^{-1})^*$, it follows from (2.3) that

$$(B')^{-1} = I_V^{-1} \circ (B^{-1})^*. \tag{2.4}$$

2.2. The optimal test space norm

We now restrict ourselves to the case where the "trial space" U is a Hilbert space with an inner product $(\cdot,\cdot)_U$ and corresponding norm $\|\cdot\|_U$. We define the *optimal test norm* on the "test space" V by

$$\|v\|_V := \sup_{u\in U} \frac{|b(u,v)|}{\|u\|_U}. \tag{2.5}$$

Since B' is a bijection, this norm generates a topology equivalent to the original topology in V (so we will have no use for the original norm on V). It is easy to see that the optimal test norm is generated by the inner product

$$(w,v)_V := b\big(R_U^{-1}B'w, v\big) \tag{2.6}$$

where $R_U : U \to U'$ is the isometric Riesz operator defined by $R_U u(\delta u) = (\delta u, u)_U$ for all $u, \delta u \in U$. Thus, we have made V into a Hilbert space. Note that by the polarization identity, the $(\cdot,\cdot)_V$-inner product is uniquely determined from the V-norm. In our examples later, this inner product will be obvious from inspection and we will not need to use (2.6) to implement it.

2.3. The optimal test functions

Now we recall the *Petrov-Galerkin scheme* of [11]. In [11], the method was presented using a general inner product on V. In contrast, here we are interested in using the specific inner product $(\cdot,\cdot)_V$ with its corresponding optimal test norm $\|\cdot\|_V$ introduced above.

Let $U_N \subset U$ be a finite-dimensional space with a basis $\{e_j : j = 1,\dots,N\}$. Define $T : U \to V$ by

$$(Tu, v)_V = b(u,v), \quad \forall v \in V. \tag{2.7}$$

For each *trial* basis function e_j, the corresponding *optimal test (basis) function* is $Te_j \in V$. They form the optimal discrete test space

$$V_N := \operatorname{span}\{Te_j : j = 1,\dots,N\} \subset V. \tag{2.8}$$

The *Petrov-Galerkin scheme* for (2.2) is as follows.

$$\begin{cases} \text{Find } u_N \in U_N \text{ such that :} \\ b(u_N, v_N) = l(v_N), \quad \forall v_N \in V_N. \end{cases} \tag{2.9}$$

This can be thought of as a least square method [7] as explained in [11]. It is proven in [11, Theorem 2.2] (and the proof is a simple consequence of Babuška's theorem) that in the *energy norm* defined by

$$\|u\|_E := \sup_{v\in V} \frac{|b(u,v)|}{\|v\|_V}, \tag{2.10}$$

the solution of the *Petrov-Galerkin scheme* (2.9) is the best approximation, i.e.,

$$\|u - u_N\|_E = \inf_{w_N\in U_N} \|u - w_N\|_E. \tag{2.11}$$

While it is difficult to characterize $\|\cdot\|_E$ in general [11], because we used the optimal test norm $\|\cdot\|_V$ of § 2.2 in (2.10), we have the following simple characterization.

Proposition 2.1. *For all $u \in U$, we have $\|u\|_E = \|u\|_U$, and consequently,*

$$\|u - u_N\|_U = \inf_{w_N\in U_N} \|u - w_N\|_U. \tag{2.12}$$

Proof. By (2.5), $|b(u,v)| \leq \|u\|_U \|v\|_V$, so obviously $\|u\|_E \leq \|u\|_U$. The reverse inequality obviously follows if we prove that

$$\inf_{u\in U} \frac{\|u\|_E}{\|u\|_U} = \inf_{u\in U} \sup_{v\in V} \frac{|b(u,v)|}{\|u\|_U \|v\|_V} = \inf_{v\in V} \sup_{u\in U} \frac{|b(u,v)|}{\|v\|_V \|u\|_U} = 1. \tag{2.13}$$

The last equality in (2.13) is obvious from (2.5). The first inf-sup equals $1/\|B^{-1}\|$, while the second equals $1/\|(B')^{-1}\|$ (where $\|\cdot\|$ denotes the appropriate operator norms). They are equal by (2.4). Hence (2.13) follows and we have proved that $\|u\|_E = \|u\|_U$. Using this, (2.12) follows from (2.11). □

We note two properties of the *Petrov-Galerkin scheme* (2.9).

1. The global stiffness matrix of the method is *Hermitian and positive definite* irrespective of the symmetry properties of $b(\cdot,\cdot)$. Indeed,

   $$b(e_i, Te_j) = (Te_i, Te_j)_V = \overline{(Te_j, Te_i)}_V = \overline{b(e_j, Te_i)},$$

   so it is Hermitian. Positive definiteness follows from (2.7). This property is a manifestation of the least square nature of the method.

2. Once the approximate solution has been determined, the norm of the finite element error $e_N := u - u_N$ can be computed once we solve the following problem:

   $$\begin{cases} \text{Find } Te_N \in V \text{ such that :} \\ (Te_N, \delta v)_V = b(u - u_N, \delta v) = l(\delta v) - b(u_N, \delta v), \quad \forall \delta v \in V. \end{cases} \tag{2.14}$$

   Then,

   $$\|e_N\|_U = \|e_N\|_E = \|Te_N\|_V. \tag{2.15}$$

   We call the solution Te_N to problem (2.14) *the error representation function*. Notice that we can compute a good approximation to energy norm of the error *without knowing* the exact solution by solving an approximate version of (2.14). Indeed, the energy norm of the error is nothing other than a properly defined norm of the residual.

2.4. Equivalent test norms

In our examples later, the optimal norm $\|\cdot\|_V$ turns out to be inconvenient for practical computations. Hence we investigate changes that result when it is substituted with another norm $\|\cdot\|_{\tilde{V}}$ satisfying

$$C_1 \|v\|_{\tilde{V}} \leq \|v\|_V \leq C_2 \|v\|_{\tilde{V}}, \quad \forall v \in V. \tag{2.16}$$

We assume that the new norm is generated by a computable inner product $(\cdot,\cdot)_{\tilde{V}}$. When this is used in place of $(\cdot,\cdot)_V$ in (2.7), different optimal test functions, and consequently a different *Petrov-Galerkin scheme* results. Let us denote its solution by $\tilde{u}_N$. It is the best approximation in the following energy norm

$$\|u\|_{\tilde{E}} := \sup_{v\in V} \frac{|b(u,v)|}{\|v\|_{\tilde{V}}} \tag{2.17}$$

which in general is not equal to $\|u\|_U$. Yet, we have the following theorem.

Theorem 2.1. *Let C_1 and C_2 be the constants of the equivalence relation* (2.16). *Then*

$$\|u - \tilde{u}_N\|_U \leq \frac{C_2}{C_1} \inf_{w_N \in U_N} \|u - w_N\|_U.$$

Proof. The solution $\tilde{u}_N$ (due to the result of [11] recalled in (2.11)) satisfies

$$\|u - \tilde{u}_N\|_{\tilde{E}} = \inf_{w_N \in U_N} \|u - w_N\|_{\tilde{E}}.$$

Hence the theorem will follow if we show that

$$C_1\|u\|_U \leq \|u\|_{\tilde{E}} \leq C_2\|u\|_U, \quad \forall u \in U. \tag{2.18}$$

For any nonzero v in V, taking reciprocals in (2.16) we obtain

$$\frac{1}{C_2\|v\|_{\tilde{V}}} \leq \frac{1}{\|v\|_V} \leq \frac{1}{C_1\|v\|_{\tilde{V}}}. \tag{2.19}$$

Hence (2.18) follows by multiplying the inequality (2.19) by $|b(u, v)|$ and taking the supremum over all non-zero vectors $v \in V$. □

From Theorem 2.1, it is clear that in order to achieve the good stability in $\|\cdot\|_U$, independent of problem parameters (e.g., the wavenumber k in the problems we shall consider in the later sections), the alternative norm $\|\cdot\|_{\tilde{V}}$ should be designed so that the ratio of the equivalence factors, C_2/C_1, is (i) as small as possible, and (ii) independent of the parameters.

2.5. Practicalities.

For conforming discretizations, the application of T, which requires the solution of (2.7), leads to a global system of equations. Then the computation of the optimal test space is too expensive and the entire discussed concept has little practical value.

The situation changes if the methodology is applied in the framework of *discontinuous* Petrov-Galerkin (DPG) method. When functions in V are discontinuous across mesh elements, and when the V-inner product is locally computable, then the solution of (2.7) becomes a local operation. By approximating these local problems suitably, test functions close to optimality can be inexpensively computed. In the methods we present, we use richer polynomial spaces (a few degrees higher than the trial spaces) to approximate (2.7).

To ensure that functions in V are discontinuous, we treat all equations of a boundary value problem weakly. The starting point in the design of a DPG method is a reformulation of the boundary value problem into a system of first-order differential equations. Introducing a partitioning of the spatial domain Ω into mesh elements $\{K\}$, the equations are multiplied element-wise by test functions in a "broken" test space

$$V = V_{\text{DPG}} = \prod_K V(K), \tag{2.20}$$

integrated over the whole domain, and then integrated by parts in each element. The resulting boundary flux terms are treated rather as *independent unknowns*. Fluxes known from boundary conditions are replaced or moved to the right-hand side where appropriate, contributing in that manner to the linear functional $l(v)$. Contrary to classical variational formulations where some of the equations are relaxed and others are treated in a strong form, in the DPG method *all equations* are treated in a weak sense. Formulations like this are sometimes referred to as *ultra-weak variational formulations*.

When using a general inner product on V, we can choose it to be local (as we did in [11, 12]). Unfortunately, the optimal V-inner product of the optimal test norm is generally not

local because of the introduction of fluxes in the DPG formulation. In this situation, we must find an equivalent *localizable* $\tilde{V}$-inner product, i.e., its associated *localizable* $\tilde{V}$-norm satisfies

$$\|v\|_{\tilde{V}}^2 = \sum_K \|v_K\|_{\tilde{V}}^2, \quad \forall v \in V, \tag{2.21}$$

where v_K denotes the restriction of v to K, extended by zero to Ω. This maintains the locality of test space computations. Changes in the solution due to the substitution of the new inner product can be analyzed using the results in § 2.4.

For adaptivity, we can use the error indicator shown in (2.15). Its square equals the sum of corresponding element contributions, i.e.,

$$\|e_N\|_E^2 = \|Te_N\|_{V_{\text{DPG}}}^2 = \sum_K e_K, \tag{2.22}$$

where $e_K = \|(Te_N)_K\|_{\tilde{V}}^2$. The error representation function $\left(Te_N\right)_K$ on K is computed by solving a local counterpart of (2.14) using an element enriched space. The element contributions e_K serve as element error indicators in an *hp*-adaptive algorithm (see [12]).

3. A model time-harmonic transport problem

As a prelude to the DPG formulation of the full Helmholtz equation presented in Section 4, in this section we study a simplified 1D time-harmonic wave propagation problem. We begin by considering the *spectral* method, in which we use a globally-conforming test space V, and identify the explicit forms of the optimal test norm and inner product. We then consider the *DPG* method, utilizing a "broken" test space V_{DPG}. After identifying the optimal test norm and inner product in this setting, we present an equivalent *localizable* norm which may be practically utilized in the DPG setting. The final *hp* method is presented in § 3.3. In § 3.4, we then discuss the approximation of the optimal test functions in the numerical implementation, and present results demonstrating the stated robustness in both the spectral and DPG settings.

We consider the problem:

$$\begin{cases} ik\rho + \rho' &= 0 \quad \text{in } (0,1), \\ \rho(0) &= \rho_0. \end{cases} \tag{3.23}$$

This arises by assuming the time-dependence of the form $e^{+i\omega t}$ in a transport equation. The exact solution to (3.23) is the right-traveling plane wave

$$\rho(x) = \rho_0 e^{-ikx}.$$

3.1. Purely spectral formulation

The spectral case is the easiest to describe because we can work with "non-broken" spaces and there is just one unknown flux. Set the test space $V = H^1(0,1)$. The spectral variational formulation associated with (3.23) is

$$\begin{cases} \text{Find } (\rho, \hat{\rho}) \in U := L^2(0,1) \times \mathbb{C}, \text{ such that :} \\ \underbrace{\int_0^1 -\rho(\overline{ikq + q'}) + \hat{\rho}\overline{q(1)}}_{b((\rho,\hat{\rho}),q)} = \rho_0 \overline{q(0)}, \quad \forall q \in V := H^1(0,1). \end{cases} \tag{3.24}$$

Note the presence of the flux unknown $\hat{\rho}$.

As we saw in section 2.2, the choice of the norm in the trial space U determines the optimal norm and inner product on the test space V. We choose

$$\|(\rho,\hat{\rho})\|_U^2 := \|\rho\|_{L^2(0,1)}^2 + |\hat{\rho}|^2,$$

as we would like the discrete solution to converge in L^2. The optimal test norm defined by (2.5) is then

$$\|q\|_V = \sup_{(\rho,\hat{\rho})\in U} \frac{|b((\rho,\hat{\rho}),q)|}{\|(\rho,\hat{\rho})\|_U}.$$

This supremum is immediately verified to be

$$\|q\|_V^2 = \|ikq + q'\|_{L^2(0,1)}^2 + |q(1)|^2. \tag{3.25}$$

The inner product which generates this norm is also obvious:

$$(q,\delta q)_V = \left(ikq + q', ik\delta q + \delta q'\right)_{L^2(0,1)} + q(1)\overline{\delta q(1)}. \tag{3.26}$$

Next, we pick a trial space discretization. To obtain a spectral method, we can simply set

$$U_N \equiv U_p := P_p(0,1)\times\mathbb{C}$$

where $P_p(0,1)$ denotes the space of (complex) polynomials of degree at most p on $(0,1)$. Then $V_N \equiv V_p$ is obtained as in (2.8) once we specify what T is for this example. For each e in U_N, the function $q = Te$ in V solves:

$$\begin{cases} \text{Find } q \equiv Te \in V \text{ such that:} \\ \\ (ikq + q', ik\delta q + \delta q')_{L^2(0,1)} + q(1)\overline{\delta q(1)} = b(e,\delta q) \quad \forall \delta q \in V. \end{cases} \tag{3.27}$$

With U_p and V_p thus defined, our spectral DPG approximation of 3.23, namely $(\rho_p,\hat{\rho}_p)\in U_p$, is obtained by solving (2.9). By Proposition 2.1,

$$\begin{aligned} \|\rho-\rho_p\|_{L^2(0,1)}^2 + |\hat{\rho}-\hat{\rho}_p|^2 &= \inf_{(w_p,\hat{w}_p)\in U_p} \|\rho - w_p\|_{L^2(0,1)}^2 + |\hat{\rho}-\hat{w}_p|^2 \\ &= \inf_{w_p\in U_p} \|\rho - w_p\|_{L^2(0,1)}^2, \end{aligned} \tag{3.28}$$

i.e., the discrete solution ρ_p coincides with the $L^2(0,1)$-orthogonal projection of the exact solution ρ into the polynomial space. Moreover, $\hat{\rho} = \hat{\rho}_p$.

3.2. An intermediate method

We now modify the above spectral method in two steps: First, we set U_N to a discrete space of discontinuous functions based on a mesh (since the trial space of the formulation (3.24) is based in L^2, this results in minimal modifications). Partition $(0,1)$ into n elements using

$$0 = x_0 < x_1 < \ldots < x_{j-1} < x_j < \ldots < x_n = 1. \tag{3.29}$$

Setting polynomial degrees p_j on each element $K_j = (x_{j-1}, x_j)$, define

$$\begin{aligned} L^2_{hp} &:= \{w : w|_{K_j} \in P_{p_j}(K_j)\}, \\ U_N \equiv \breve{U}_{hp} &:= L^2_{hp}\times\mathbb{C}. \end{aligned} \tag{3.30}$$

The second modification is to change the inner product on V from (3.26) to

$$(q,\delta q)_{\breve{V}} = (ikq + q', ik\delta q + \delta q')_{L^2(0,1)} + \frac{1}{2}(q,\delta q)_{L^2(0,1)}. \tag{3.31}$$

Lemma 3.1. *The norm $\|q\|_{\tilde{V}}$ generated by the above inner product is equivalent to the optimal V-norm in* (3.25)*, i.e.,* (2.16) *holds for all v in $H^1(0,1)$ with*

$$C_1 = (2-\sqrt{2})^{\frac{1}{2}}, \qquad C_2 = (2+\sqrt{2})^{\frac{1}{2}}.$$

The proofs of all lemmas, including this, can be found in Appendix A. The reason for considering the modified $\tilde{V}$-inner product will be clear in § 3.3. We say that q is a *global optimal test function* for problem (3.23) if it is the optimal test function corresponding to some $(\rho,\hat{\rho}) \in \breve{U}_{hp}$ in this setting, i.e.,

$$q \in H^1(0,1): \qquad (q,\delta q)_{\tilde{V}} = \int_0^1 -\rho(\overline{ik\delta q + \delta q'}) + \hat{\rho}\overline{\delta q(1)}, \quad \forall \delta q \in H^1(0,1). \tag{3.32}$$

We set V_N to $\breve{V}_{hp}$, the span of all such global optimal test functions corresponding to all $(\rho,\hat{\rho}) \in \breve{U}_{hp}$. We define an "intermediate method" for theoretical purposes, as follows:

$$\begin{cases} \text{Find } (\breve{\rho}_{hp},\hat{\rho}_{hp}) \in \breve{U}_{hp} \text{ such that :} \\ \\ \displaystyle\int_0^1 -\breve{\rho}_{hp}(\overline{ikq+q'}) + \hat{\rho}\overline{q(1)} = \rho_0\overline{q_1(0)}, \quad \forall q \in \breve{V}_{hp}. \end{cases} \tag{3.33}$$

Theorem 3.1. *We have the error estimate (with wave number independent constant)*

$$\|\rho - \breve{\rho}_{hp}\|_{L^2(0,1)} \leq \left(\frac{2+\sqrt{2}}{2-\sqrt{2}}\right)^{\frac{1}{2}} \inf_{w_{hp}\in L^2_{hp}} \|\rho - w_{hp}\|_{L^2(0,1)}.$$

Proof. Apply Theorem 2.1 (its assumption is verified by Lemma 3.1). □

3.3. The DPG method

The difficulty with the above defined intermediate method is that the computation of the optimal test space by (3.32) is a global problem, due to the global H^1-conformity of the test space. To move to a more practical method, we now "break" the test space using the mesh (3.29), namely set

$$V = V_{\text{DPG}} = \prod_{j=1}^n V(K_j), \qquad V(K_j) := H^1(x_{j-1},x_j),$$

i.e., the test functions now have the form $q = (q_1,\ldots,q_n) \in V_{\text{DPG}}$, where $q_j \in H^1(x_{j-1},x_j)$. With this we can now state the DPG variational formulation of (3.23):

$$\begin{cases} \text{Find } (\rho,\hat{\rho}) \in U := L^2(0,1)\times\mathbb{C}^n \text{ such that :} \\ \\ \underbrace{\displaystyle\sum_{j=1}^n \int_{x_{j-1}}^{x_j} -\rho(\overline{ikq_j + q_j'}) + \hat{\rho}_j\overline{[q]}_j}_{b((\rho,\hat{\rho}),q)} = \rho_0\overline{q_1(0)}, \quad \forall q \in V_{\text{DPG}}, \end{cases} \tag{3.34}$$

where we have introduced the jumps defined by

$$[q]_j = \begin{cases} q_j(x_j) - q_{j+1}(x_j) & \text{if } j = 1,...,n-1, \\ q_n(1) & \text{if } j = n, \end{cases}$$

and the vector $\hat{\rho} = (\hat{\rho}_1, \ldots, \hat{\rho}_n)$ of fluxes at element interfaces.

We choose the following norm on U:

$$\|(\rho, \hat{\rho})\|_U^2 = \|\rho\|_{L^2(0,1)}^2 + \sum_{j=1}^{n} |\hat{\rho}_j|^2.$$

Then, similar to the spectral case, an explicit expression of the optimal test norm is easily found:

$$\|q\|_{V_{\mathrm{DPG}}}^2 = \sum_{j=1}^{n} \|ikq_j + q_j'\|_{L^2(x_{j-1},x_j)}^2 + |[q]_j|^2 . \tag{3.35}$$

The above norm is induced by the inner product

$$(q, \delta q)_{V_{\mathrm{DPG}}} = \sum_{j=1}^{n} (ikq_j + q_j', ik\delta q_j + \delta q_j')_{L^2(x_{j-1},x_j)} + [q]_j \overline{[\delta q]_j}.$$

Obviously, the norm above does not satisfy the localization property (2.21). We wish to replace it with a norm that does, in order to locally compute optimal test functions and obtain local error indicators. We use the norm (and associated inner product) given by:

$$\begin{aligned} \|q\|_{\tilde{V}}^2 &= \textstyle\sum_{j=1}^{n} \|ikq_j + q_j'\|_{L^2(K_j)}^2 + \frac{1}{2}\|q_j\|_{L^2(K_j)}^2 \\ (q, \delta q)_{\tilde{V}} &= \textstyle\sum_{j=1}^{n} \Big(ikq_j + q_j', ik\delta q_j + \delta q_j'\Big)_{L^2(K_j)} + \frac{1}{2}(q_j, \delta q_j)_{L^2(K_j)}. \end{aligned} \tag{3.36}$$

Note that this is the same norm as in (3.31), when applied to q in $H^1(0,1)$.

The optimal test space is computed with the above $\tilde{V}$-inner product and the following discrete trial space

$$U_{hp} = L_{hp}^2 \times \mathbb{C}^n \subset U$$

where L_{hp}^2 is as defined in (3.30). Let $\{\rho_\ell\}$ denote the a basis for L_{hp}^2 consisting of functions each of which are supported only one element. Then a basis for U_{hp} takes the form $(\rho_\ell, \hat{e}_m)$ where $\{\hat{e}_1, \ldots, \hat{e}_n\}$ denote the standard unit basis for $\mathbb{C}^n$. The corresponding optimal test functions can now be computed locally (unlike (3.32)), so we call them the *local optimal test functions*. If ρ_ℓ is supported on K_j, then the local optimal test function q for the trial basis $(\rho_\ell, 0)$ is supported solely on K_j and is computed by solving

$$\big(ikq + q', ik\delta q + \delta q'\big)_{L^2(K_j)} + \frac{1}{2}(q, \delta q)_{L^2(K_j)} = \int_{x_{j-1}}^{x_j} -\rho_\ell(\overline{ik\delta q + \delta q'}), \quad \forall \delta q \in V(K_j).$$

Similarly, the local optimal test function corresponding to the trial basis $(0, \hat{e}_j)$ is supported on $K_j \cup K_{j+1}$ and is obtained by solving

$$\begin{aligned} \big(ikq_j + q_j', ik\delta q_j + \delta q_j'\big)_{L^2(K_j)} + \frac{1}{2}(q_j, \delta q_j)_{L^2(K_j)} &= \delta q_j(x_j), \\ \big(ikq_{j+1} + q_{j+1}', ik\delta q_{j+1} + \delta q_{j+1}'\big)_{L^2(K_{j+1})} + \frac{1}{2}(q_{j+1}, \delta q_{j+1})_{L^2(K_{j+1})} &= -\delta q_{j+1}(x_j), \end{aligned}$$

for all δq in V_{DPG}. We set the test space V_N to V_{hp}, the span of these optimal test functions. Clearly $V_{hp} \subseteq V_{\mathrm{DPG}}$. The DPG method is then given as follows.

$$\begin{cases} \text{Find } (\rho_{hp}, \hat{\rho}^{hp}) \in U_{hp} \text{ such that :} \\ \\ \displaystyle\sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} -\rho_{hp}(\overline{ikq_j + q_j'}) + \hat{\rho}_j^{hp}\overline{[q]}_j = \rho_0\overline{q_1(0)}, \quad \forall q \in V_{hp}. \end{cases} \tag{3.37}$$

We analyze this method using the following lemma, proved in Appendix A.

Lemma 3.2. *The global optimal test functions in $\breve{V}_{hp}$ are contained in V_{hp}. Consequently the solutions $\breve{\rho}_{hp}$ of (3.33) and ρ_{hp} of (3.37) coincide:*

$$\breve{\rho}_{hp} = \rho_{hp}.$$

Theorem 3.2. *An error estimate holds with a constant independent of wave number:*

$$\|\rho - \rho_{hp}\|_{L^2(0,1)} \leq \left(\frac{2+\sqrt{2}}{2-\sqrt{2}}\right)^{\frac{1}{2}} \inf_{w_{hp} \in L^2_{hp}} \|\rho - w_{hp}\|_{L^2(0,1)}. \tag{3.38}$$

Proof. By Lemma 3.2, $\rho_{hp} = \breve{\rho}_{hp}$, so (3.38) follows immediately from Theorem 3.1. □

3.4. Numerical results

As noted in Section 2.5, we rely on higher order approximation in order to approximate the optimal test functions spanning the discrete test space V_{hp}. More precisely, corresponding to our trial space U_{hp}, we form an enriched test space $V^+_{hp} \subset V$ from which we approximate the optimal test functions. Given an element $K_j = (x_{j-1}, x_j)$ of polynomial order p_j, its discrete enriched local test space $V^+_{hp}(K_j) \subset H^1(x_{j-1}, x_j)$ is taken to be $P_{p_j+\Delta p}(K_j)$, where the parameter $\Delta p \geq 1$ is the *degree of enrichment*. The local test space $V_{hp}(K_j) \subset V(K_j)$ is determined by solving the discrete local problems

$$\begin{cases} \text{Find } q_j^{hp} \in V^+_{hp}(K_j) \text{ such that:} \\ \\ (q_j^{hp}, \delta q)_{V(K_j)} = b_{K_j}(e, \delta q) \quad \forall \delta q \in V^+_{hp}(K_j) \end{cases}$$

for each trial basis function $e \in U_{hp}$ supported in element K_j. Here $b_{K_j}(\cdot,\cdot)$ denotes the localized sesquilinear form, defined for $q_j \in V(K_j)$ by

$$b_{K_j}((\rho,\hat{\rho}), q_j) := \begin{cases} \displaystyle\int_K -\rho(\overline{ikq_j + q_j'}) + \hat{\rho}_j q_j(x_j) - \hat{\rho}_{j-1} q_j(x_{j-1}) & j > 1, \\ \\ \displaystyle\int_K -\rho(\overline{ikq_j + q_j'}) + \hat{\rho}_j q_j(x_j) & j = 1. \end{cases} \tag{3.39}$$

In all of the following examples, we take $\rho_0 = 1$ in (3.23). We begin with the one element case, where there is no distinction between the spectral and DPG methods. Given an element of order p, the order $p + \Delta p$ of the enriched discrete test space V^+_{hp} is taken using $\Delta p = 6$. As the plots of the solutions and the measured L^2 errors indicate (Figure 1), this is sufficient to realize practically perfect L^2 stability for the considered wavenumbers; the L^2 projected and DPG solutions are visually indistinguishable, and the L^2 error of the DPG solution is within 10^{-6} percent of the projection.

Lowering the degree of enrichment Δp results in a gradual degradation in performance until the dimensions of the trial and enriched space are equal (Fig. 2). Higher enrichment obviously achieves better approximation of the optimal test functions, at the cost of more computational effort - we utilize a standard Cholesky factorization routine to solve the local system of equations, so the effort scales roughly as $(p+\Delta p)^3$. However, we emphasize that this is a local, element-by-element operation, and therefore the resources required for computation of optimal test functions are expected to be negligible compared to the cost of solving the final, global system. Moreover,

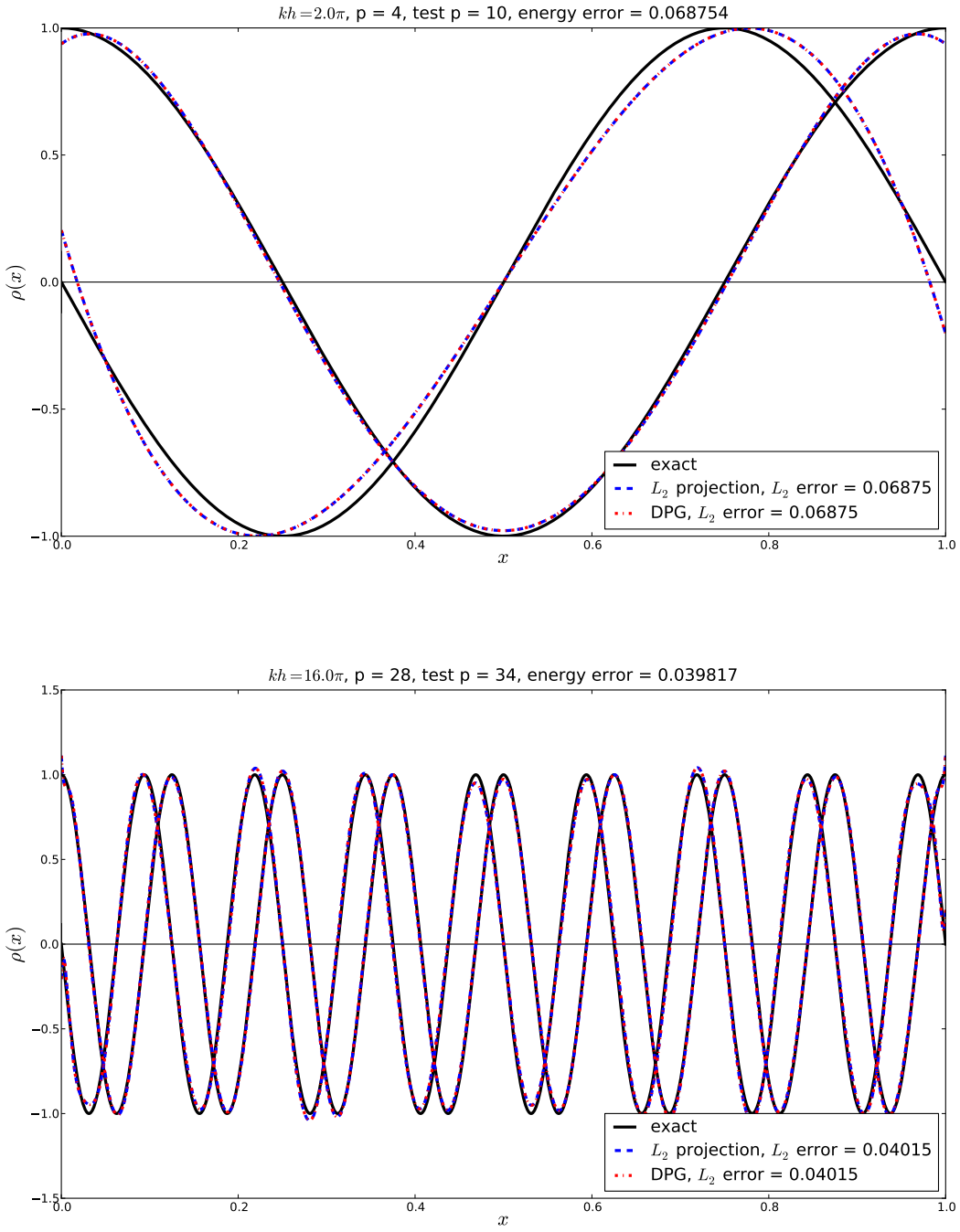


Figure 1: Comparison of the exact solution to the model problem (3.23) with the DPG solution and the best L^2 approximation, at two different wavenumbers ($k = 2\pi$ and 16π, respectively). Both real and imaginary components of the solution ρ are shown. Discretizations employed in both examples consist of one high-order element. Optimal test functions are computed corresponding to the optimal test space inner product (3.26), with degree of enrichment $\Delta p = 6$. In both examples, the DPG solution coincides with the best approximation.

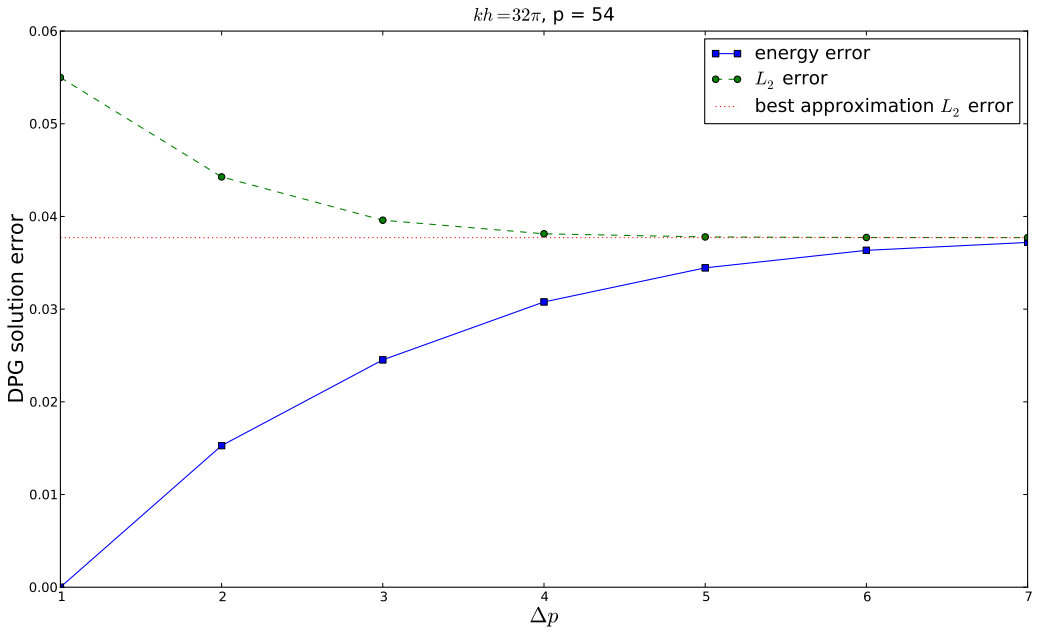


Figure 2: Here we show the effect of the degree of enrichment Δp on the error of the solution by fixing the wavenumber k and discrete trial space, and solving the problem with varying Δp. The optimal test functions are computed using the optimal test inner product (3.26). More enrichment provides better approximations to the true optimal test functions, resulting in less L^2 error. The reported energy error is computed by solving a discrete version of (2.14) over the enriched test space. Higher enrichment also brings this estimate closer to the true energy error defined by (2.15).

in many applications we are likely to encounter redundant local problems (e.g. a patch of uniform elements with identical material parameters), so optimal test functions computed on one element may be cached and re-used.

For the DPG setting, we employ the localizable inner product (3.36). In Figure 3, we plot solutions obtained with four linear elements per wavelength, for k corresponding to one and 16 wavelengths over the length of the domain, respectively. We observe very good L^2 stability, as indicated by the ratio of the DPG error to the best approximation error, regardless of wavenumber, as illustrated in Fig. 4.

4. The Helmholtz model problem

In this section, we study the coupled Helmholtz problem in 1D, represented in terms of *pressure* p and *velocity* u, coupled in a system of first-order differential equations. Again, we demonstrate that with a proper choice of the norm for the test functions, the stability properties turn out to be *wavenumber-independent*. We start by analyzing the spectral problem and we prove robustness of the method in this simple setting. Then we introduce the DPG formulation that delivers the same FE solution as the spectral one, when using the same trial FE basis. Hence, the robustness result will apply to the DPG solution as well.

4.1. The variational equations

We consider the Helmholtz equation written as a system of two first order equations. The new unknowns have physical meaning, e.g., in the theory of acoustical disturbances [8]. Given an inflow data $u_0 \in \mathbb{C}$, the speed of sound c, and the density of the fluid ρ, we consider the boundary value problem

$$\begin{cases} ik\dfrac{p}{c\rho} + u' &= 0 \qquad \text{in } \Omega = (0,1), \\ ik\, c\rho\, u + p' &= 0 \qquad \text{in } \Omega = (0,1), \\ u(0) &= u_0 \qquad \text{and} \\ p(1) &= Zu(1), \end{cases} \tag{4.40}$$

where Z is an impedance parameter relating p and u at the right side boundary. The choice $Z = c\rho$ leads to the exact right-traveling wave solution

$$u(x) = u_0 e^{-ikx}, \quad p(x) = c\rho u(x).$$

For the sake of simplicity, we will just consider the values $Z = c\rho = 1$. Note that we use the notation p for both polynomial degree and pressure. Which is meant is amply evident from the context, so no confusion will arise.

In the context of an *unbroken* test space, the natural variational formulation associated with (4.40) is

$$\begin{cases} \text{Find } (p, u, \hat{p}_0, \hat{u}_1) \in U := [L^2(\Omega)]^2 \times \mathbb{C}^2, \text{ such that :} \\ \\ ik\displaystyle\int_\Omega p\overline{q} - \int_\Omega u\overline{q}' + \hat{u}_1\overline{q}(1) = u_0\overline{q}(0), \\ \\ ik\displaystyle\int_\Omega u\overline{v} - \int_\Omega p\overline{v}' + \hat{u}_1\overline{v}(1) - \hat{p}_0\overline{v}(0) = 0, \quad \forall (q,v) \in V := [H^1(\Omega)]^2. \end{cases} \tag{4.41}$$

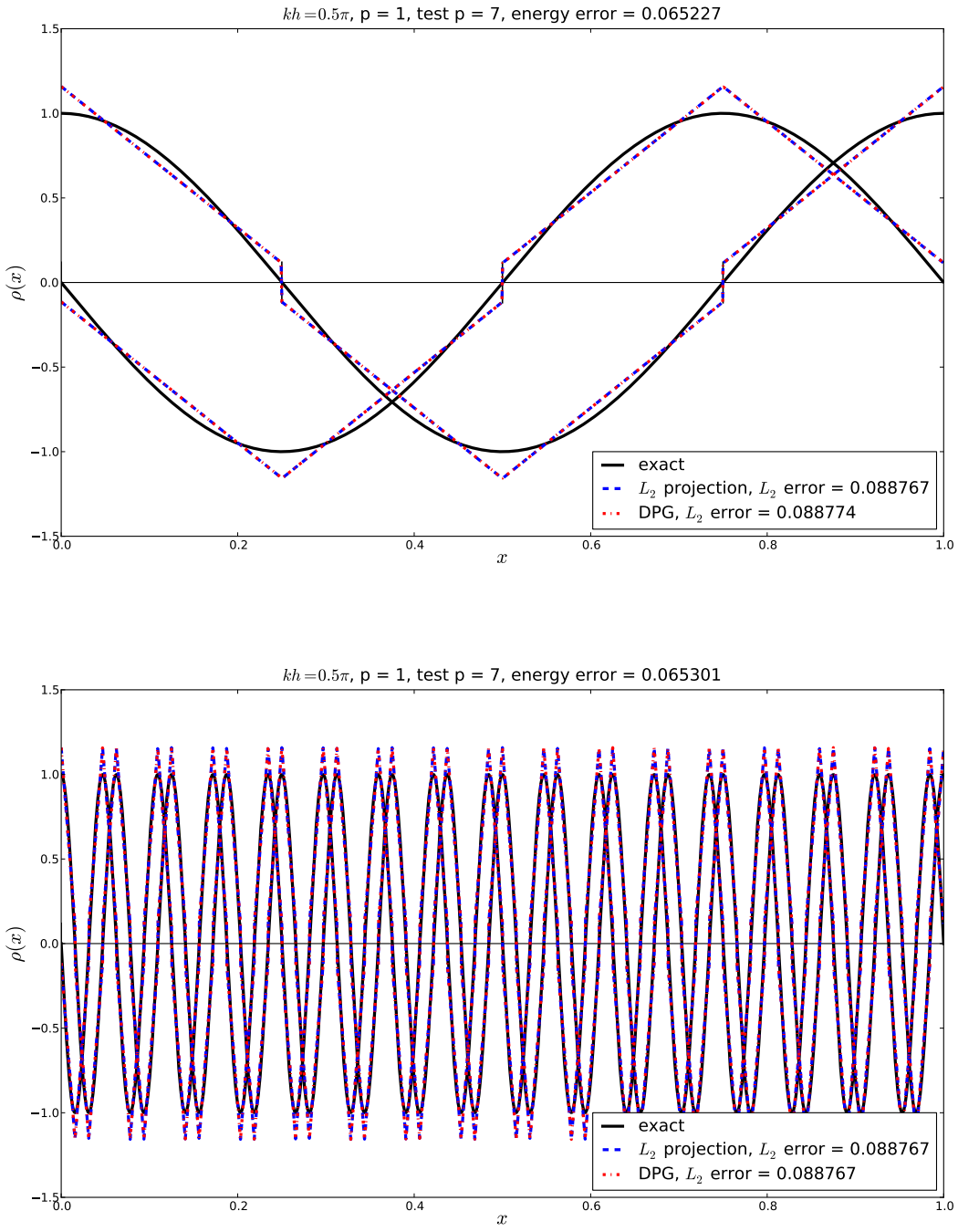


Figure 3: Comparison of the exact solution to the model problem (3.23) with the DPG solution and the best L^2 approximation, at two different wavenumbers ($k = 2\pi$ and 32π, respectively). Both real and imaginary components of the solution ρ are shown. Discretizations employed in both examples consist of four linear elements per wavelength. Optimal test functions are computed corresponding to the localizable inner product (3.31), with degree of enrichment $\Delta p = 6$. In both examples, the DPG solution nearly coincides with the best approximation, even when k is increased several times in the bottom example.

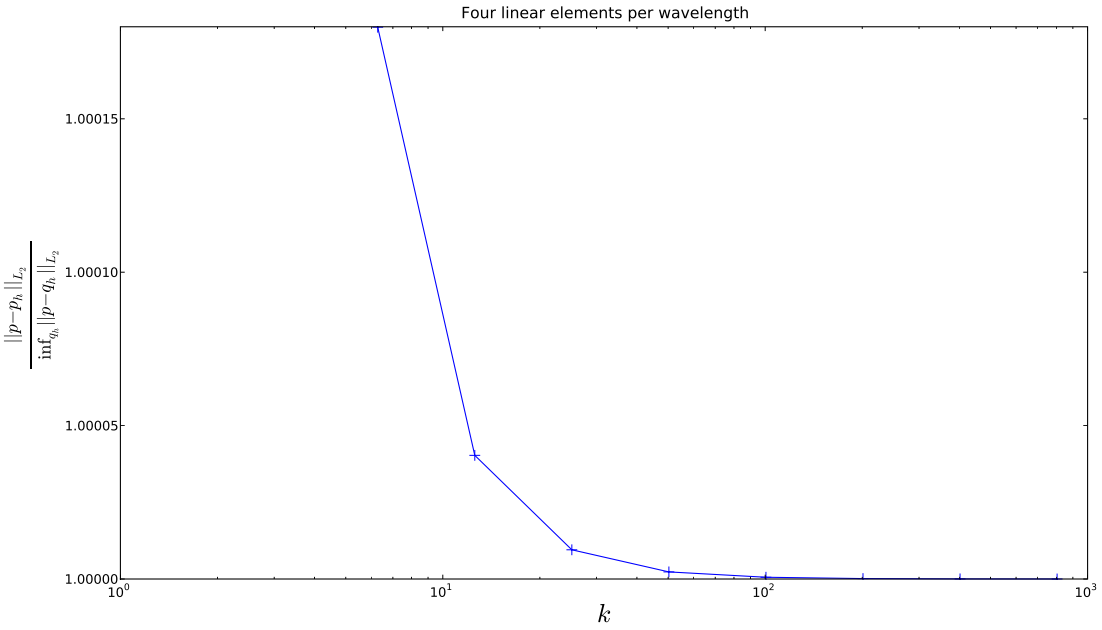


Figure 4: As indicated by Theorem 3.2, the DPG method employing the localizable inner product (3.31) is robust with respect to wavenumber k. Here we take discretizations of four linear elements per wavelength and plot the ratio of the DPG error to the best approximation error as the wavenumber is increased, observing that the ratio approaches a wavenumber-independent value (which is also very close to one).

As in § 3.1, this will immediately lead to a spectral method, once we select the trial space norm in which we would like convergence, i.e.,

$$\|p, u, \hat{p}_0, \hat{u}_1\|_U^2 := \|p\|_{L^2}^2 + \|u\|_{L^2}^2 + |\hat{p}_0|^2 + |\hat{u}_1|^2.$$

Then, the sesquilinear form on the left side of (4.41) yields the following *optimal inner product and norm* on V:

$$\begin{aligned}\big((q,v),(\delta q,\delta v)\big)_V = &\ \big(ikq+v', ik\delta q+\delta v'\big)_{L^2} + \big(ikv+q', ik\delta v+\delta q'\big)_{L^2} \\ &+\big(q(1)+v(1)\big)\big(\overline{\delta q}(1)+\overline{\delta v}(1)\big) + v(0)\overline{\delta v}(0) \qquad (4.42)\\ \|q,v\|_V^2 = &\ \|ikq+v'\|_{L^2}^2 + \|ikv+q'\|_{L^2}^2 + \big|q(1)+v(1)\big|^2 + |v(0)|^2.\end{aligned}$$

Inspired by what we have observed in the previous section, we replace (4.42) with the following localizable norm (and associated inner product):

$$\begin{aligned}\|q,v\|_{\tilde{V}}^2 &= \|ikq+v'\|_{L^2}^2 + \|ikv+q'\|_{L^2}^2 + \tfrac{1}{\alpha}\|v+q\|_{L^2}^2 + \tfrac{1}{\alpha}\|v\cos(kx)+iq\sin(kx)\|_{L^2}^2\\ \big((q,v),(\delta q,\delta v)\big)_{\tilde{V}} &= (ikv+q', ik\delta v+\delta q')_{L^2} + (ikq+v', ik\delta q+\delta v')_{L^2} + \tfrac{1}{\alpha}(v+q,\delta v+\delta q)_{L^2}\\ &\quad +\tfrac{1}{\alpha}\big(v\cos(kx)+iq\sin(kx), \delta v\cos(kx)+i\delta q\sin(kx)\big)_{L^2},\end{aligned} \qquad (4.43)$$

where $\alpha = \frac{5+\sqrt{5}}{2}$. Then we have the norm equivalence result (2.16), as stated in the next lemma. Its proof is in A.

Lemma 4.1. *For all (q,v) in V,*

$$C_1\|q,v\|_{\tilde{V}}^2 \le \|q,v\|_V^2 \le C_2\|q,v\|_{\tilde{V}}^2,$$

with constants $C_1 = \frac{5+\sqrt{5}}{2} - (5+2\sqrt{5})^{\frac{1}{2}}$ and $C_2 = \frac{5+\sqrt{5}}{2} + (5+2\sqrt{5})^{\frac{1}{2}}$.

The next step is to consider an "intermediate method" as in § 3.2. Again let (3.29) define the mesh of $(0,1)$. The trial space in this case is $\breve{U}_{hp} = [L^2_{hp}(\Omega)]^2 \times \mathbb{C}^2$. Let $\{(p_l, u_l, \hat{p}_{0l}, \hat{u}_{1l})\}_{l=1,\dots,N} \subset [L^2(\Omega)]^2 \times \mathbb{C}^2$ be a finite trial basis such that each p_l and u_l are in L^2_{hp} and are supported on a single element. With these as the trial functions and (4.43) as the inner product, we find the corresponding optimal test functions. These are the *global optimal test functions* for this case and we denote their span by $\breve{V}_{hp}$. Let $\breve{\mathfrak{U}}_{hp} = (p_{hp}, u_{hp}, \hat{p}_{hp}, \hat{u}_{hp})$ be the discrete solution from this method. Then we have the following robust error estimate for the intermediate problem.

Theorem 4.1. *Let $\breve{\mathfrak{U}} = (p, u, \hat{p}, \hat{u})$ be the exact solution of problem (4.41). Then*

$$\|\breve{\mathfrak{U}} - \breve{\mathfrak{U}}_{hp}\|_U \le \left(\sqrt{\frac{5+\sqrt{5}}{2}} + \sqrt{\frac{3+\sqrt{5}}{2}}\right) \inf_{\mathcal{W}_{hp}\in\breve{U}_{hp}} \|\breve{\mathfrak{U}} - \mathcal{W}_{hp}\|_U. \qquad (4.44)$$

Proof. This follows from Theorem 2.1 and Lemma 4.1. □

As before, the global optimal test functions are expensive to compute. Hence we formulate the DPG method with local optimal test functions next.

4.2. The DPG formulation

Using the same partition (3.29) of the interval $(0,1)$, we formulate the DPG method as follows. The unknowns include field variables (p,u) and fluxes $\hat{p}_{j-1}, \hat{u}_j$, $j=1,\dots,n$. Fluxes $\hat{u}_0$, $\hat{p}_n$ were replaced by u_0 and $\hat{u}_n$ respectively by using the boundary conditions. For each element (x_{j-1},x_j), we consider local test functions $(q_j,v_j) \in [H^1(x_{j-1},x_j)]^2$. The DPG variational formulation is :

$$\begin{cases} \text{Find } (p,u,\hat{p},\hat{u}) \in L^2(\Omega)\times L^2(\Omega)\times \mathbb{C}^n \times \mathbb{C}^n \text{ such that :} \\ \\ \displaystyle\sum_{j=1}^{n} ik\int_{x_{j-1}}^{x_j} p\overline{q}_j - \int_{x_{j-1}}^{x_j} u\overline{q}_j' + (\hat{u}\overline{q})\Big|_{x_{j-1}}^{x_j} = 0 \\ \displaystyle\sum_{j=1}^{n} ik\int_{x_{j-1}}^{x_j} u\overline{v}_j - \int_{x_{j-1}}^{x_j} p\overline{v}_j' + (\hat{p}\overline{v})\Big|_{x_{j-1}}^{x_j} = 0, \quad \forall (q_j,v_j) \in [H^1(K_j)]^2. \end{cases} \tag{4.45}$$

Recall that $\hat{u}_0 = u_0$ is known and is moved to the right-hand side. Similarly, $\hat{p}_n$ is replaced by $\hat{u}_n$ in the last term of the sum. The *broken test space* in the variational formulation (4.45) is endowed with the norm and inner product induced by (4.43):

$$\begin{aligned} \|q,v\|_{\tilde{V}}^2 = & \sum_{j=1}^{n} \|ikq_j + v_j'\|_{L^2}^2 + \|ikv_j + q_j'\|_{L^2}^2 \\ & + \frac{1}{\alpha}\sum_{j=1}^{n} \|v_j + q_j\|_{L^2}^2 + \|v_j\cos(kx) + iq_j\sin(kx)\|_{L^2}^2 \\ \big((q,v),(\delta q,\delta v)\big)_{\tilde{V}} = & \sum_{j=1}^{n} \big(ikv_j + q_j', ik\delta v_j + \delta q_j'\big)_{L^2} + \big(ikq_j + v_j', ik\delta q_j + \delta v_j'\big)_{L^2} \\ & + \frac{1}{\alpha}\sum_{j=1}^{n} (v_j + q_j, \delta v_j + \delta q_j)_{L^2} \\ & + \frac{1}{\alpha}\sum_{j=1}^{n} \big(v_j\cos(kx) + iq_j\sin(kx), \delta v_j\cos(kx) + i\delta q_j\sin(kx)\big)_{L^2}. \end{aligned}$$

The discrete trial space is obtained by substituting L^2_{hp} for $L^2(0,1)$. The corresponding optimal test functions computed using the above $\tilde{V}$-inner product gives the *local optimal test functions*. Their span gives the test space V_{hp} which defines the full DPG method. Let $(p_{hp},u_{hp},\hat{p}_{hp},\hat{u}_{hp})$ be the discrete solution.

Theorem 4.2. *The global optimal test functions are contained in the space of local optimal test functions. Consequently,*

$$\|u-u_{hp}\|_{L^2}^2 + \|p-p_{hp}\|_{L^2}^2 \le \left(\sqrt{\frac{5+\sqrt{5}}{2}} + \sqrt{\frac{3+\sqrt{5}}{2}}\right)^2 \inf_{w_{hp},s_{hp}\in L^2_{hp}} \|u-w_{hp}\|_{L^2}^2 + \|p-s_{hp}\|_{L^2}^2.$$

The proof of this theorem is similar to the proof of Theorem 3.2 (cf. proof of Lemma 3.2), so we omit it.

Based on the discussion made in Section 2.4, the robustness estimation (4.44) also applies for the FE solution of this formulation, under analog considerations of the discrete trial space, aside from the additional $2(n-1)$ fluxes.

4.3. Numerical results

Figure 5 depicts solutions obtained with the spectral method, taking $u_0 = 1$, $Z = 1$. Again we have taken $\Delta p = 6$ in constructing the enriched test space, which achieves nearly perfect L^2 stability. Only the pressures p are plotted; the error in the velocity u is practically identical.

The optimal test functions (q, v) in the spectral case may be expressed as the solutions of the problem:

$$\begin{cases} ikq + v' & = -p \quad \text{in } (0,1), \\ ikv + q' & = -u \quad \text{in } (0,1), \\ v(0) & = -\hat{p}_0, \\ q(1) + v(1) & = \hat{u}_1. \end{cases}$$

The optimal test functions for a number of basis functions are illustrated in Figures 6a-6d. Their oscillatory behavior corresponding to wavenumber k makes clear the necessity of using sufficient order $p + \Delta p$ within the enriched test space.

For the multi-element case, we employ the localizable norm (4.43). Corresponding optimal test functions are shown in Figures 7a-7d. Solutions and corresponding errors obtained with the localizable norm are shown in Fig. 8. Again, we observe excellent stability with the DPG method. In contrast to standard methods, there is no degradation in stability with increasing wavenumber; indeed, if we adhere to a "rule of thumb" of n elements per wavelength as we vary k, we observe that the stability constant converges to a k-independent value (Figures 9a-9b.) For comparison, we plot in Fig. 10 a standard H^1-conforming, Bubnov-Galerkin approximation p_{hp}, as well as an H^1-conforming approximation p_{blended} obtained using specialized quadrature rules which reduce the phase lag (see [1]). While they are obtained using the same discretization, i.e. order and number of elements, we must note that they only require solution of the pressure field p, while the DPG method requires the introduction of additional variables u, $\hat{p}$, and $\hat{u}$, requiring significantly more degrees of freedom.

Even at extremely poor discretization (e.g. 2 linear elements per wavelength), the method remains stable, delivering results very near the L^2 projection (Fig. 11).

Figure 12 shows a solution obtained using a high-order ($p = 4$) DPG method.

4.4. Helmholtz with PML truncation

Finally, we consider the Helmholtz problem with PML truncation at the right end of the domain, admitting only outgoing and evanescent waves. We employ the PML stretching function

$$z(x) = \begin{cases} x & \text{in } (0, \ell) \\ x - i\frac{C}{k}\left(\frac{x-\ell}{1-\ell}\right)^4 + \frac{C}{k}\left(\frac{x-\ell}{1-\ell}\right)^4 & \text{in } (\ell, 1), \end{cases}$$

where ℓ is the position at which the PML begins, and C is a parameter controlling the strength of the PML (we take $C = 745$ to achieve decay to near machine epsilon (double precision) for a

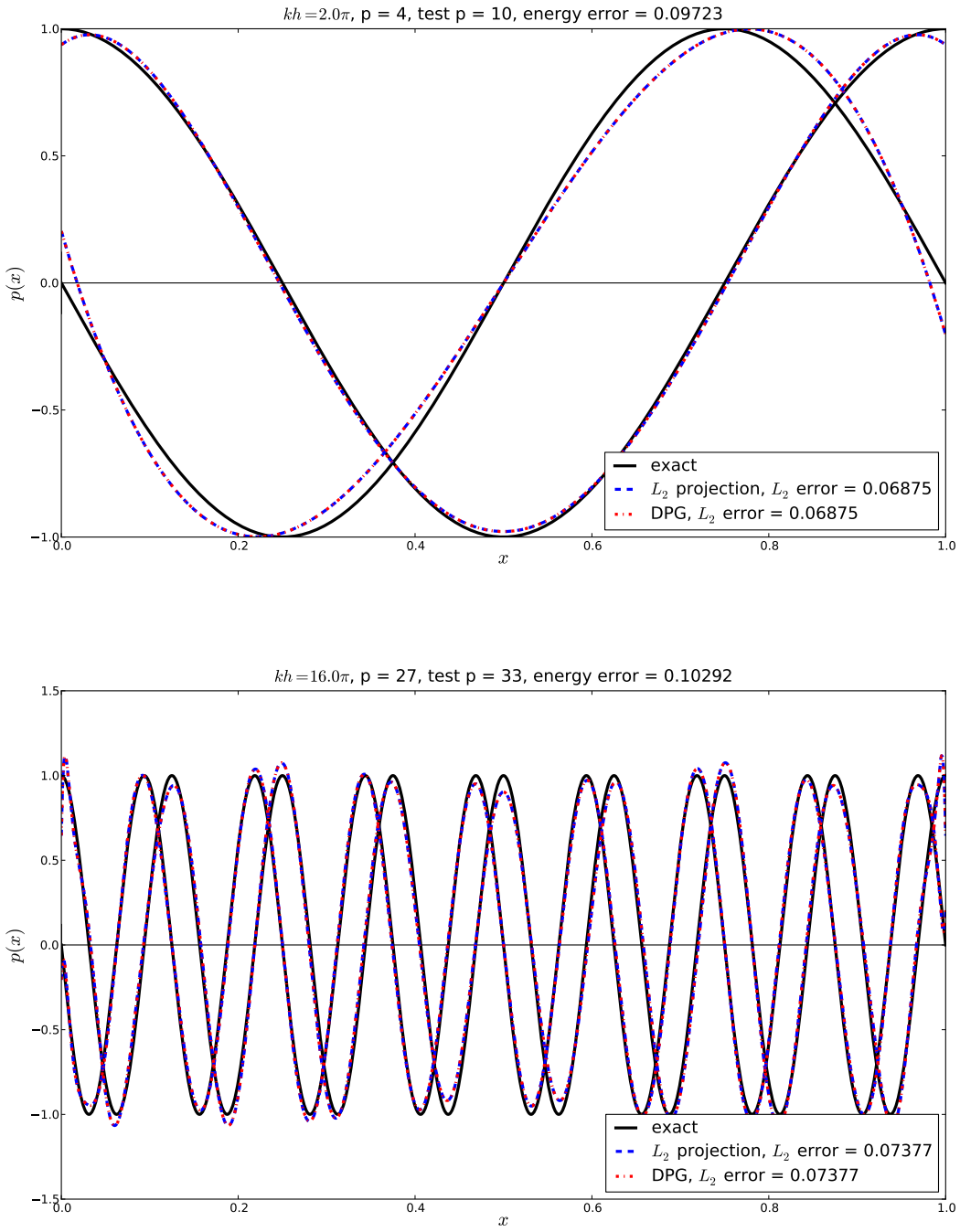
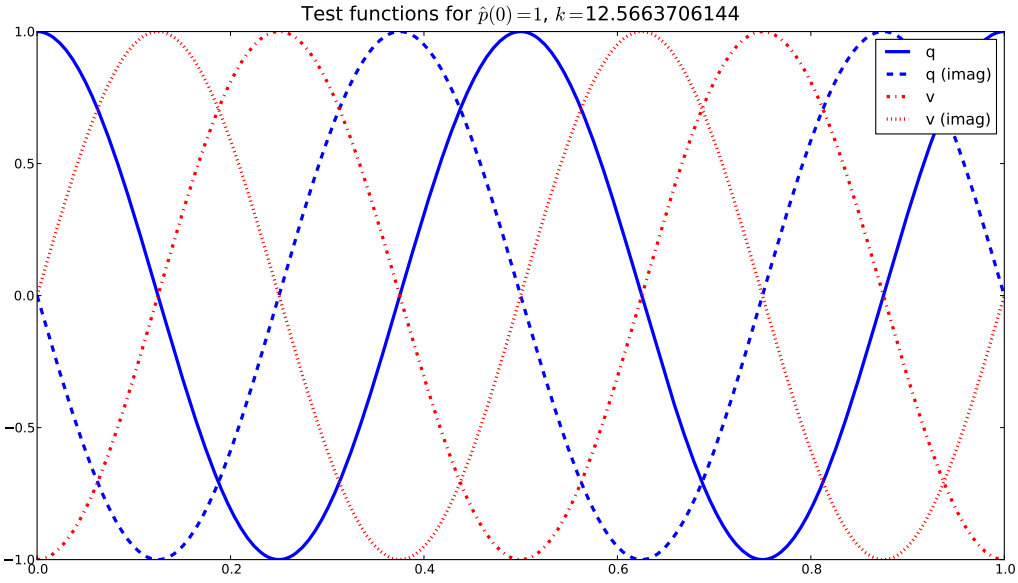
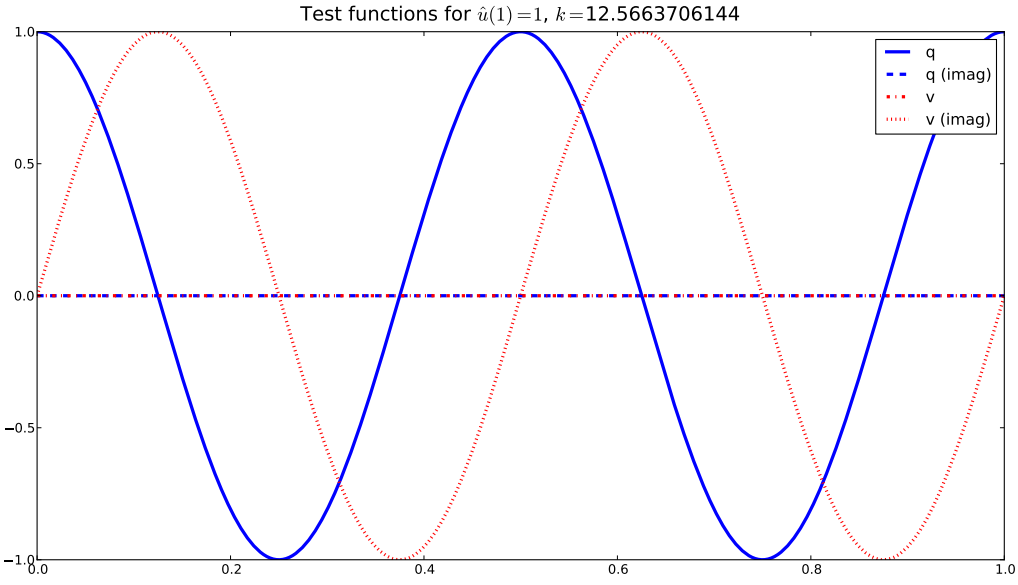


Figure 5: Comparison of the exact solution to the Helmholtz model problem (4.40) with the spectral method solution and the best L^2 approximation, at two different wavenumbers ($k = 2\pi$ and 16π, respectively). Both real and imaginary components of the pressure p are shown. Discretizations employed in both examples consist of one high-order element. Optimal test functions are computed corresponding to the optimal test space inner product (4.42), with degree of enrichment $\Delta p = 6$. In both examples, the spectral method solution coincides with the best approximation.

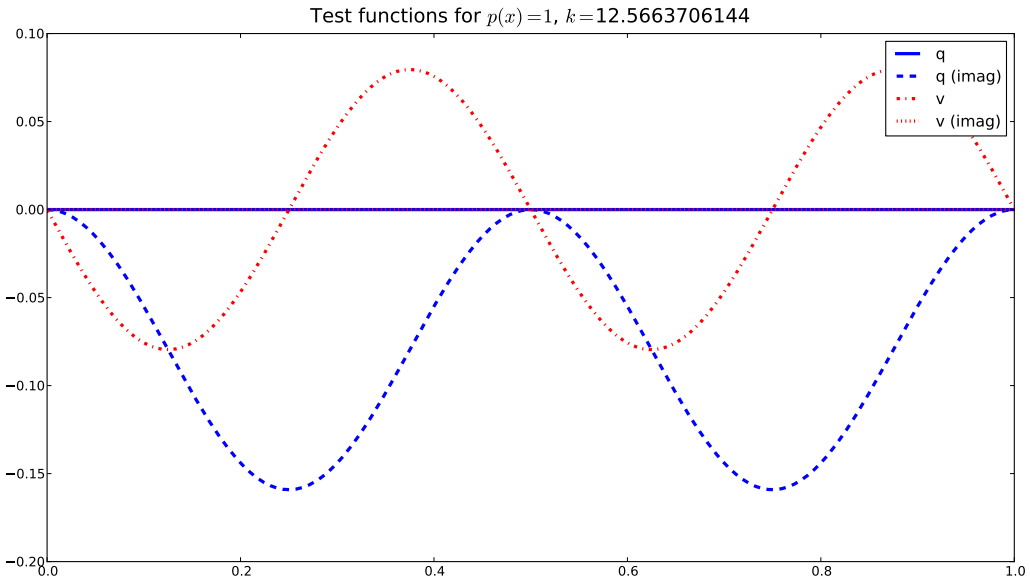


(a) Optimal test function for basis flux $\hat{p}_0 = 1$

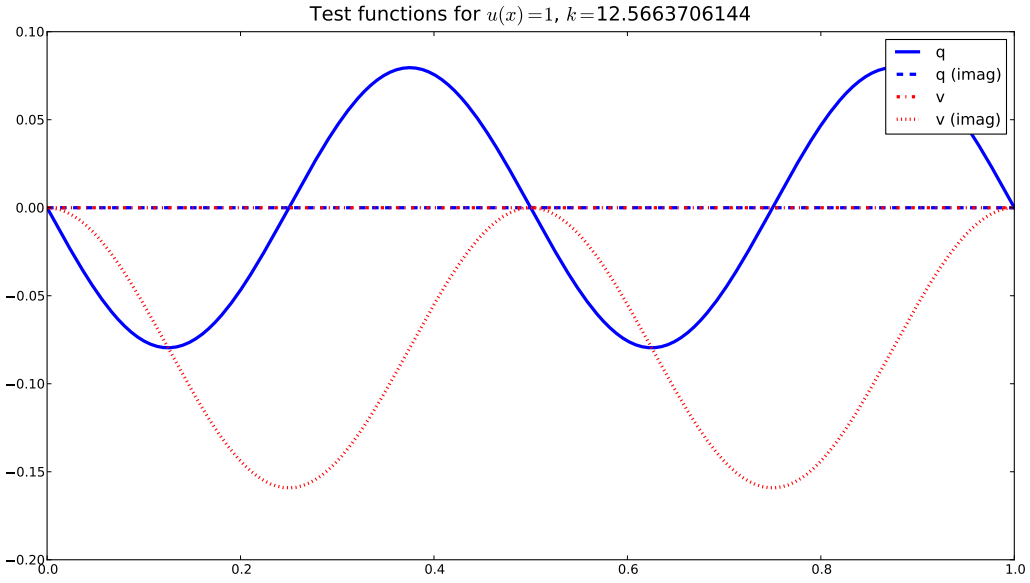


(b) Optimal test function for basis flux $\hat{u}_1 = 1$

Figure 6: Plots of the optimal test functions (q, v) for the Helmholtz spectral method, for $k = 4\pi$. The functions are defined using the optimal test inner product (4.42).

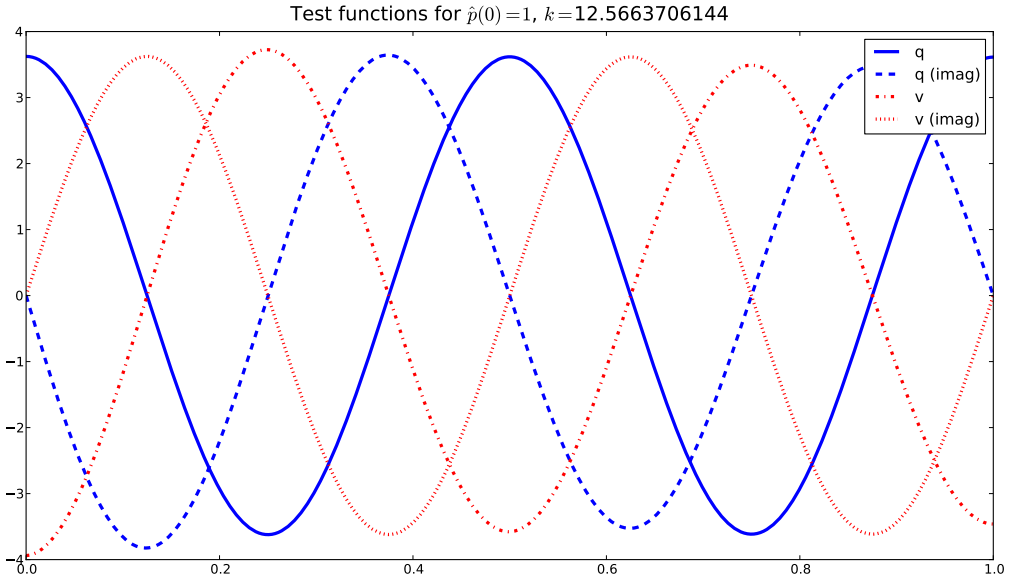


(c) Optimal test function for basis function $p(x) = 1$

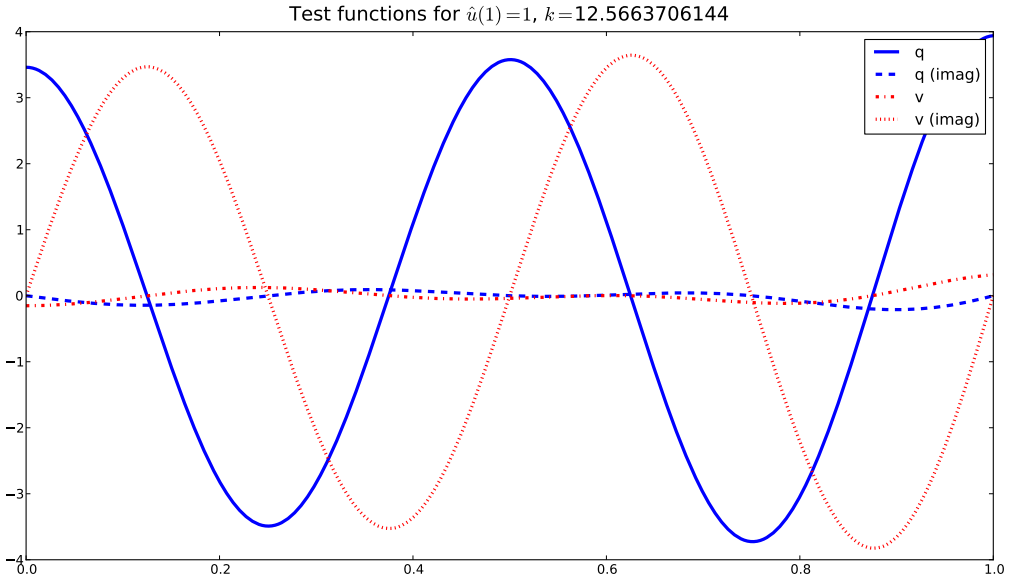


(d) Optimal test function for basis function $u(x) = 1$

Figure 6: (continued) Plots of the optimal test functions (q, v) for the Helmholtz spectral method, for $k = 4\pi$. The functions are defined using the optimal test inner product (4.42).

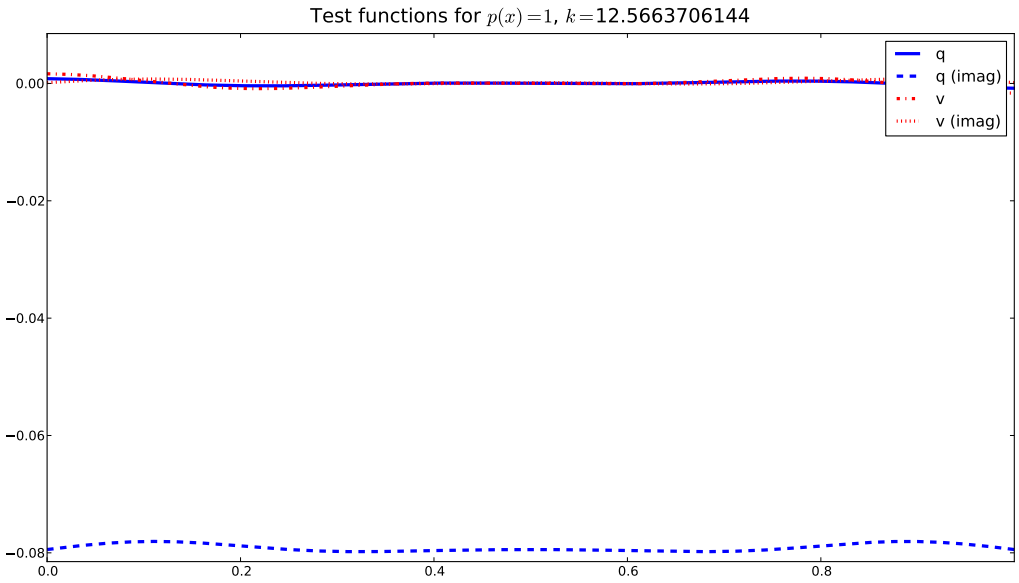


(a) Optimal test function for basis flux $\hat{p}_0 = 1$

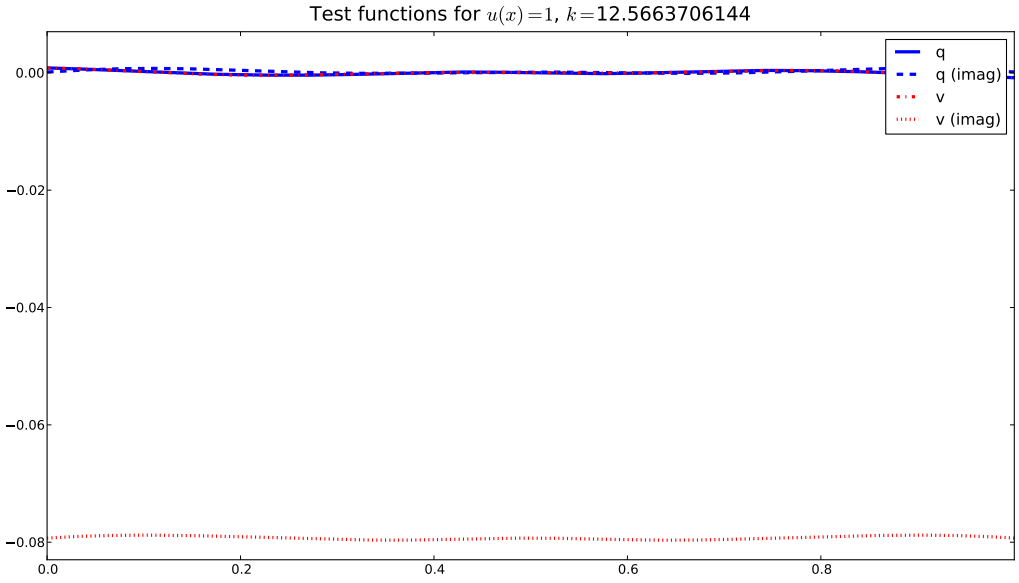


(b) Optimal test function for basis flux $\hat{u}_1 = 1$

Figure 7: Plots of the optimal test functions (q, v) for the Helmholtz spectral method, for $k = 4\pi$. The functions are defined using the localizable inner product (4.43).



(c) Optimal test function for basis function $p(x) = 1$



(d) Optimal test function for basis function $u(x) = 1$

Figure 7: (continued) Plots of the optimal test functions (q, v) for the Helmholtz spectral method, for $k = 4\pi$. The functions are defined using the localizable inner product (4.43).

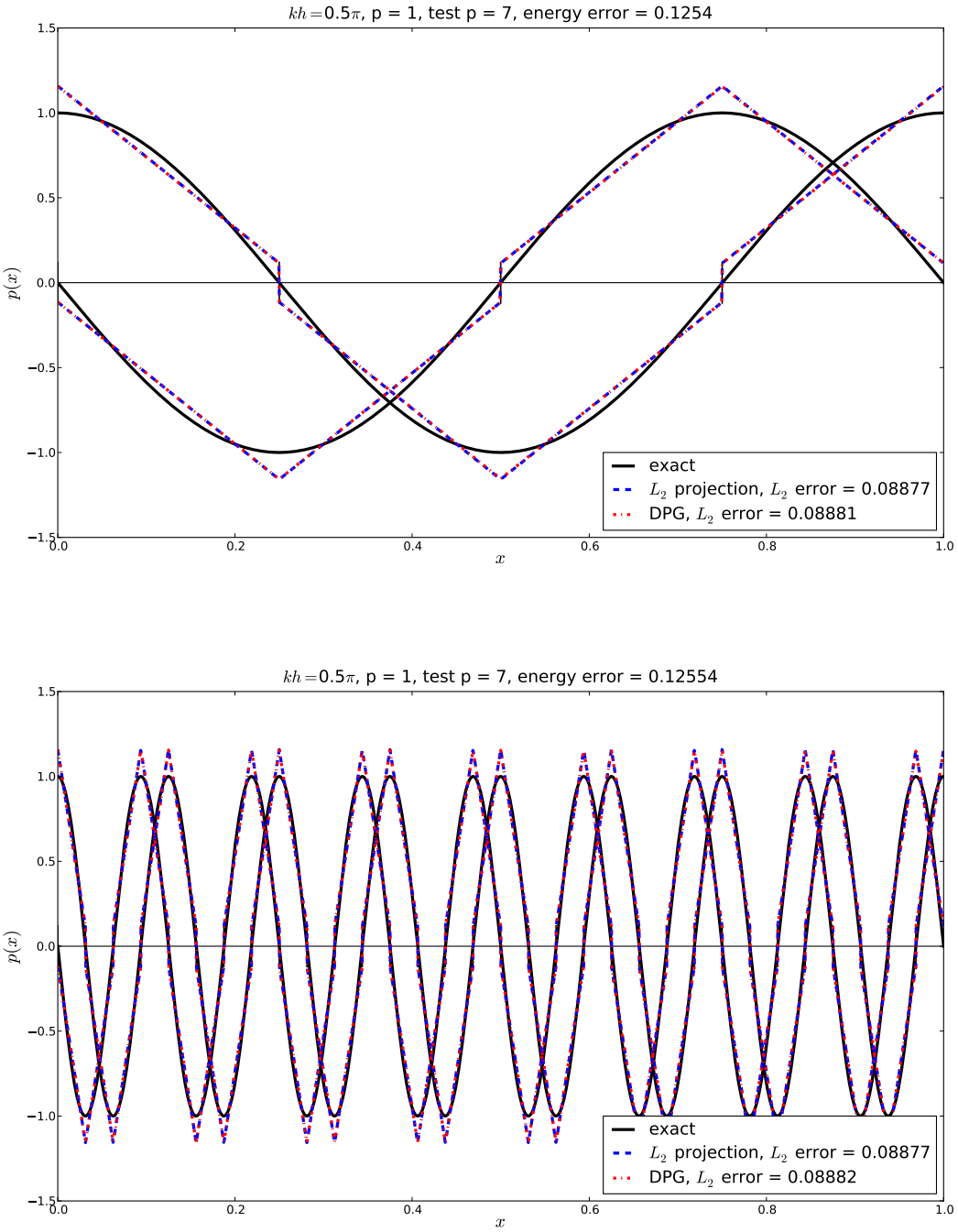
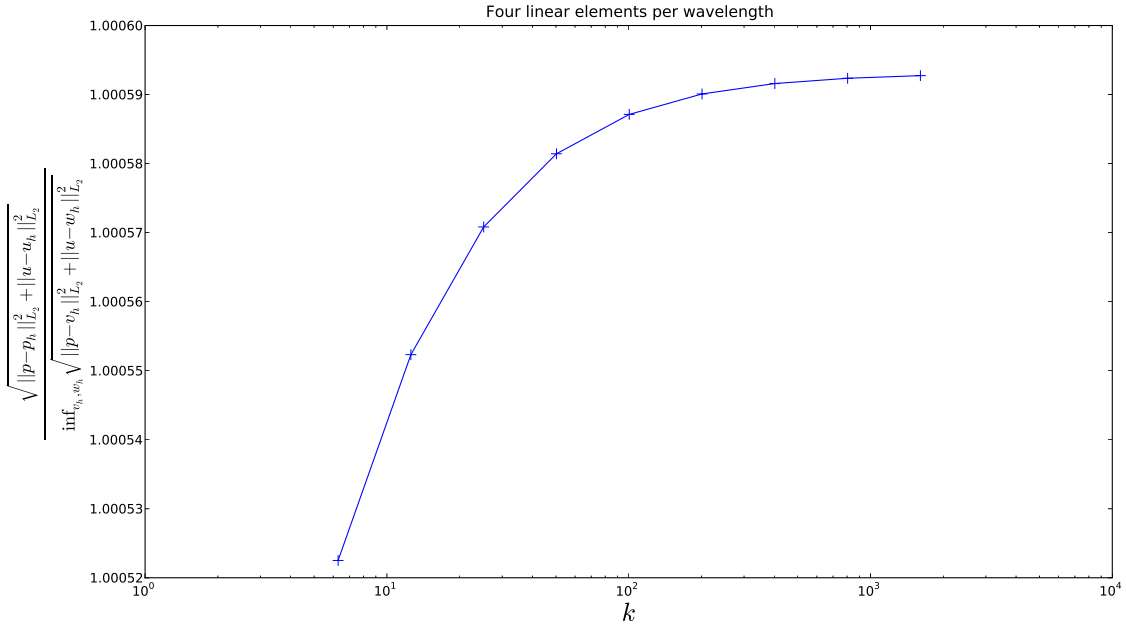
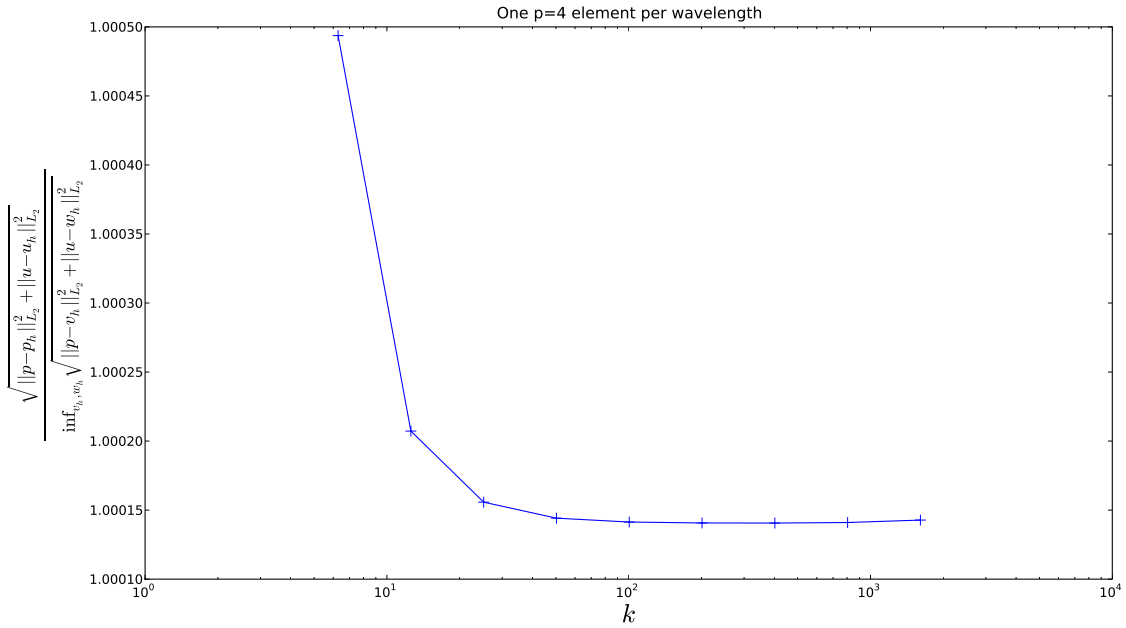


Figure 8: Comparison of the exact solution to the Helmholtz model problem (4.40) with the DPG solution and the best L^2 approximation, at two different wavenumbers ($k = 2\pi$ and 16π, respectively). Both real and imaginary components of the pressure p are shown. Discretizations employed in both examples consist of four linear elements per wavelength. Optimal test functions are computed corresponding to the localizable inner product (4.43), with degree of enrichment $\Delta p = 6$. In both examples, the DPG solution is very close to the best approximation.



(a) Four linear elements per wavelength



(b) One $p = 4$ element per wavelength

Figure 9: As indicated by Theorem 4.2, the DPG method employing the localizable inner product (4.43) is robust with respect to wavenumber k. Here we take discretizations of four linear elements per wavelength (top) and one $p = 4$ element per wavelength (bottom) and plot the ratio of the DPG error to the best approximation error as the wavenumber is increased, observing that the ratio approaches a wavenumber-independent value (which is also very close to one).

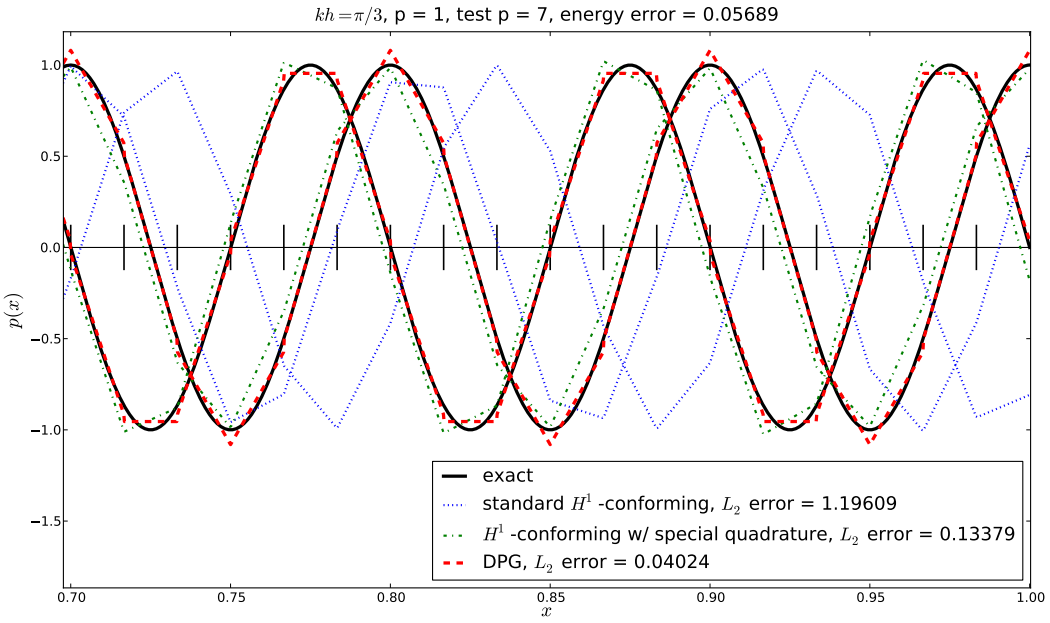


Figure 10: Comparison of the exact solution to the Helmholtz model problem (4.40) with the DPG solution, a standard H^1-conforming FEM solution, and another H^1-conforming method employing specialized quadrature rules which reduce phase error. Both real and imaginary parts of the pressure p are shown. Six linear elements per wavelength are used with each method. Optimal test functions for the DPG method are computed corresponding to the localizable inner product (4.43), with degree of enrichment $\Delta p = 6$.

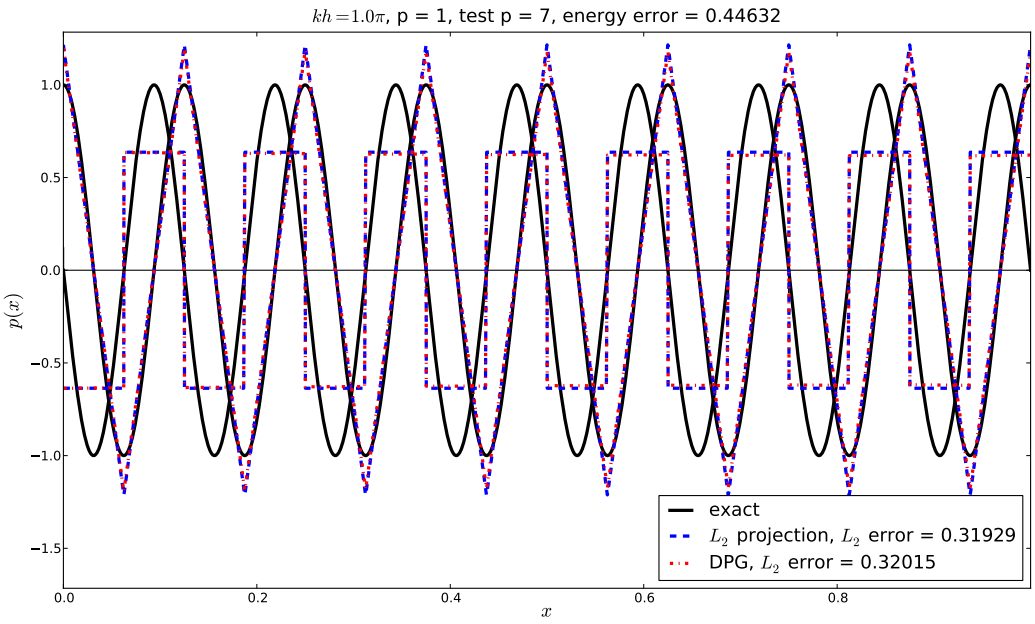


Figure 11: Comparison of the exact solution to the Helmholtz model problem (4.40) with the DPG solution and the best L^2 approximation. Both real and imaginary components of the pressure p are shown. In this example, we take a rather coarse discretization of two linear elements per wavelength, but still obtain results very near the L^2 best approximation. Optimal test functions are computed corresponding to the localizable inner product (4.43), with degree of enrichment $\Delta p = 6$.

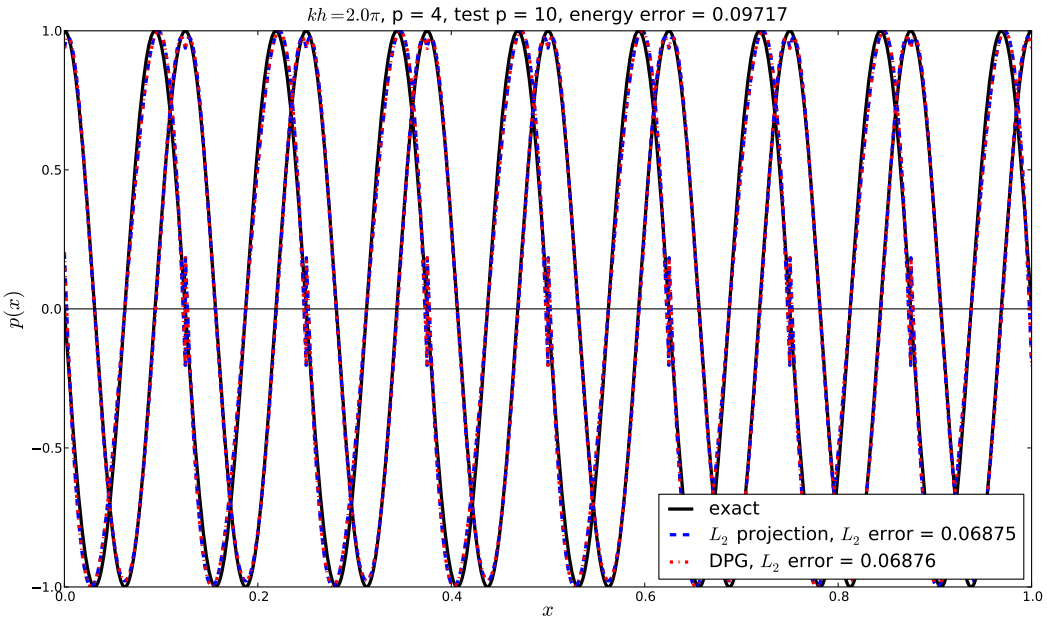


Figure 12: Comparison of the exact solution to the Helmholtz model problem (4.40) with the DPG solution and the best L^2 approximation, at $k = 16\pi$. Both real and imaginary components of the pressure p are shown. The trial space discretization consists of one $p = 4$ element per wavelength. Optimal test functions are computed corresponding to the localizable inner product (4.43), with degree of enrichment $\Delta p = 6$.

plane wave of unit amplitude). The PML problem is then:

$$\begin{cases} ikz'p + u' &= 0 \quad \text{in } \Omega = (0,1), \\ ikz'u + p' &= 0 \quad \text{in } \Omega = (0,1), \\ u(0) &= u_0, \\ p(1) &= 0. \end{cases}$$

The DPG variational formulation is:

Find $(p, u, \hat{\boldsymbol{p}}, \hat{\boldsymbol{u}}) \in L^2(0,1) \times L^2(0,1) \times \mathbb{C}^n \times \mathbb{C}^n$ such that:

$$\sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} -p(\overline{ik\overline{z}'q_j + v_j'}) - u(\overline{ik\overline{z}'v_j + q_j'}) + \hat{u}_j\overline{[\boldsymbol{q}]_j} + \hat{p}_{j-1}\overline{[\boldsymbol{v}]_{j-1}} \quad = u_0 q_1(0) \quad \forall (\boldsymbol{q}, \boldsymbol{v}) \in V_{\text{DPG}}. \tag{4.46}$$

The optimal test norm is evidently:

$$\|\boldsymbol{q}, \boldsymbol{v}\|_{V_{\text{DPG}}} = \sum_{j=1}^{n} \|ik\overline{z}'q_j + v_j'\|^2_{L^2(x_{j-1},x_j)} + \|ik\overline{z}'v_j + q_j'\|^2_{L^2(x_{j-1},x_j)} + |[\boldsymbol{q}]_j|^2 + |[\boldsymbol{v}]_{j-1}|^2 . \tag{4.47}$$

For multi-element computations, we employ an equivalent localizable norm:

$$\|\boldsymbol{q}, \boldsymbol{v}\|_{V_{\text{DPG}}} = \sum_{j=1}^{n} \|ik\overline{z}'q_j + v_j'\|^2_{L^2(x_{j-1},x_j)} + \|ik\overline{z}'v_j + q_j'\|^2_{L^2(x_{j-1},x_j)} + \|q_j\|^2_{L^2(x_{j-1},x_j)} + \|v_j\|_{L^2(x_{j-1},x_j)}. \tag{4.48}$$

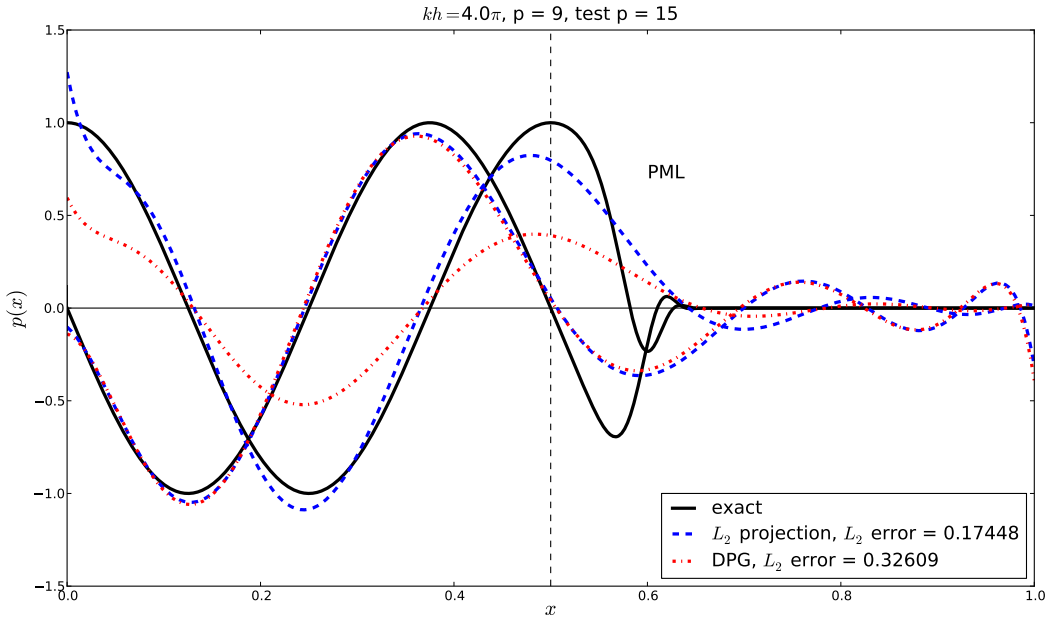
Evidently (Figures 13a-13b), when applying the spectral method to the PML truncated problem, we have to dramatically enrich our test space to realize in practice the optimal test functions delivering nearly perfect L^2 stability. This is not surprising, given that the PML stretching introduces very severe coefficients that are likely to result in wildly-behaving optimal test functions which are not well approximated using spectral methods.

In Figures 14a-14b, the PML occupies only one linear element; this is another somewhat pathological example where we must again take a very enriched test space (at least for the element occupying the PML) to obtain test functions providing good stability. Figures 15a-15b demonstrate that less enrichment is necessary when better discretization is used within the PML.

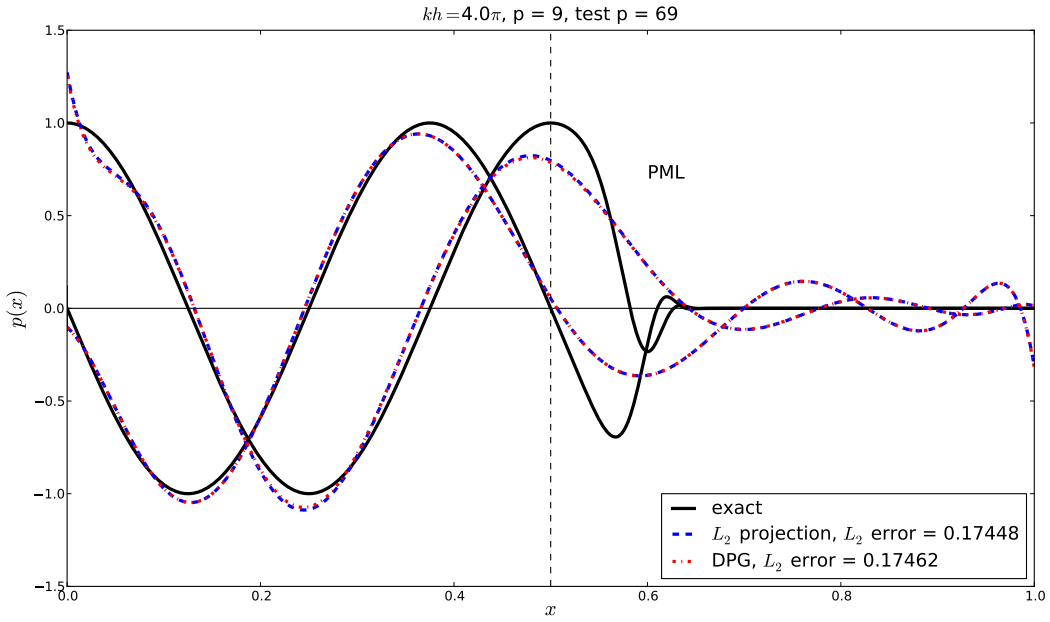
5. Conclusions

A summary.

With the introduction of the optimal test norm, an essential question of the entire DPG framework with optimal test functions has been answered: How does one design a norm on the test space V to minimize the discretization error in a given trial norm on U? The practical method additionally includes a non-trivial "norm-localization" step, in which the optimal test norm has to be replaced with a "localizable" norm, without losing uniform stability with respect to the wavenumber. The proposed methodology is very general and can apply in the multidimensional case and to other singular perturbation problems. The leading term of the equations satisfied by the optimal test functions is the L^2-adjoint of the original operator, resembling very much the old results of [13, 19, 23]. In context of singular perturbation problems, this means that the whole burden of dealing with small parameter has been moved to the problem of finding the optimal test functions. While the methodology is impractical for, e.g., problems with

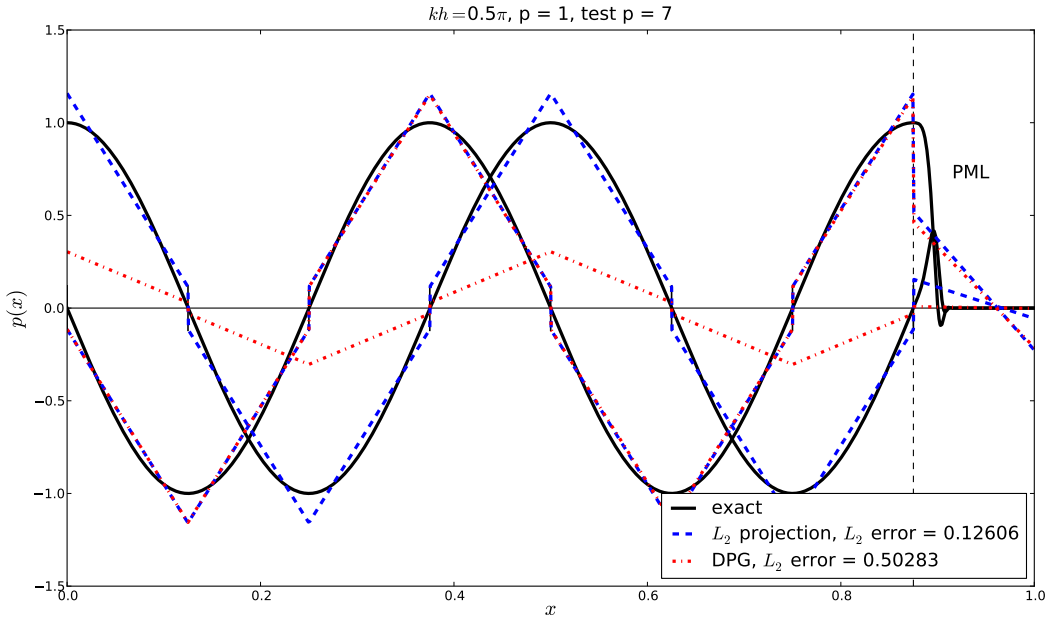


(a) Insufficiently enriched test space ($\Delta p = 6$)

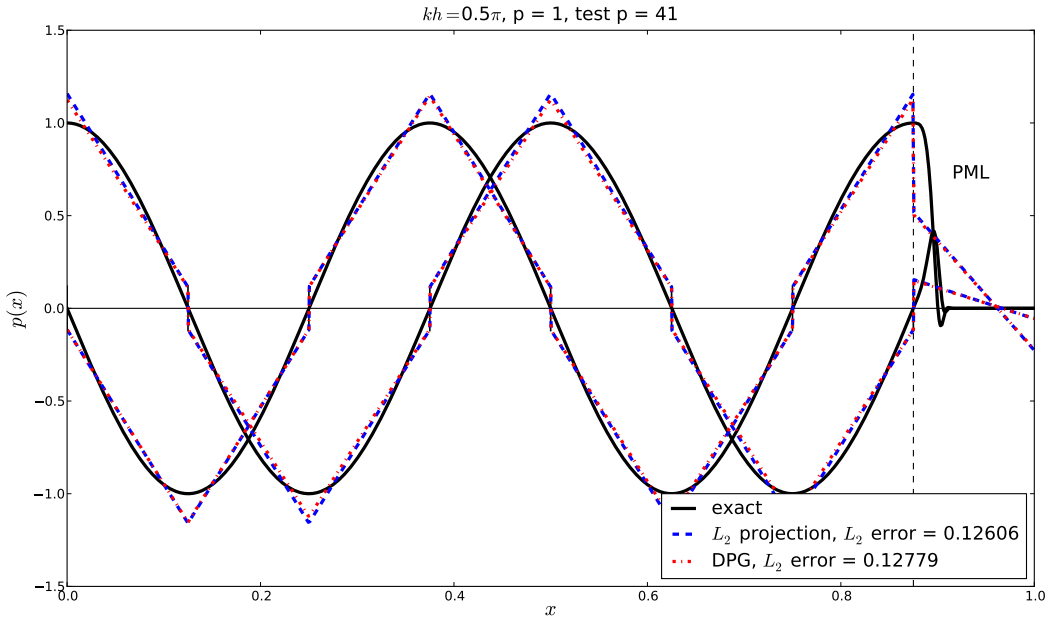


(b) Dramatically enriched test space ($\Delta p = 60$)

Figure 13: Comparison of the exact solution to the PML truncated Helmholtz problem (4.46) with the spectral method solution and the best L^2 approximation, at $k = 4\pi$. The PML truncation begins at $x = 0.5$. Both real and imaginary components of the pressure p are shown. One $p = 9$ element is employed. Optimal test functions are computed corresponding to the optimal norm (4.47); high degree of enrichment is necessary in order to approximate the true optimal test functions in this example.

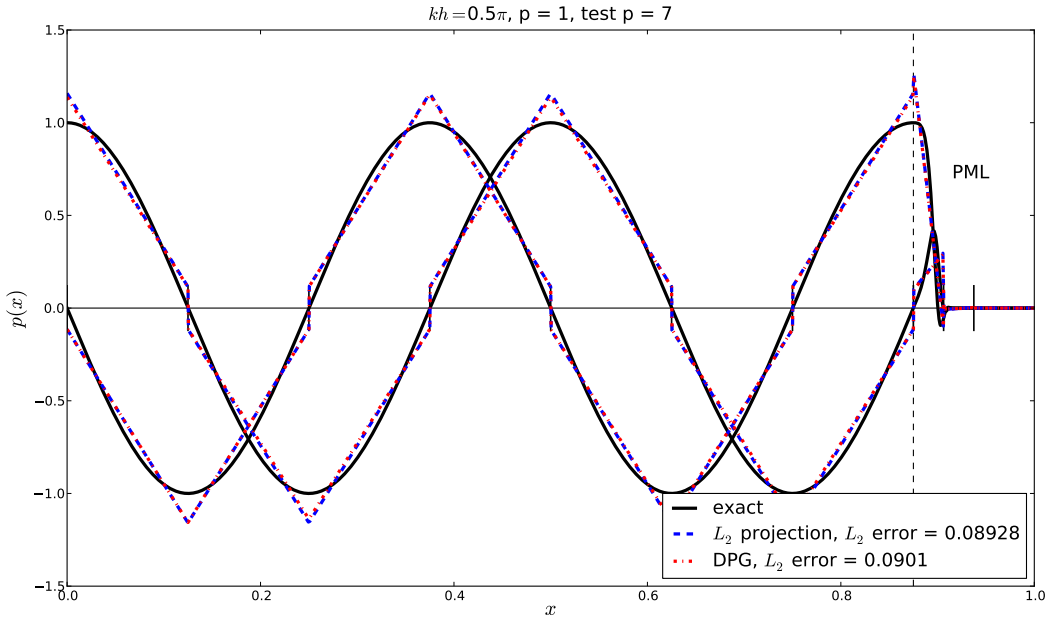


(a) Insufficiently enriched test space ($\Delta p = 6$)

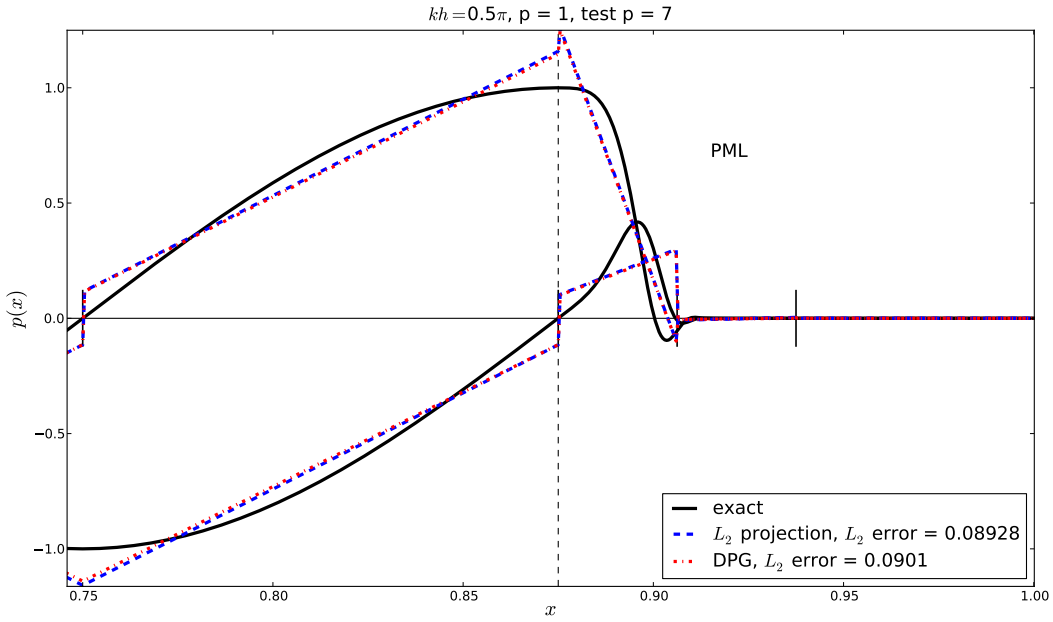


(b) Dramatically enriched test space ($\Delta p = 40$)

Figure 14: Comparison of the exact solution to the PML truncated Helmholtz problem (4.46) with the DPG solution and the best L^2 approximation. Both real and imaginary components of the pressure p are shown. Four linear elements per wavelength are used, and the PML occupies the last element. Optimal test functions are computed corresponding to the localizable norm (4.48); high degree of enrichment is necessary in order to approximate the true optimal test functions with support within the PML.



(a)



(b) Zoom of Fig. 15A near the PML

Figure 15: Comparison of the exact solution to the PML truncated Helmholtz problem (4.46) with the DPG solution and the best L^2 approximation. Both real and imaginary components of the pressure p are shown. Four linear elements per wavelength are used, except in the PML region where additional refinements have been introduced. Optimal test functions are computed corresponding to the localizable norm (4.48); with the additional refinements in the PML, we can take more reasonable degrees of enrichment and still obtain good results.

boundary layers (solving the adjoint equation on a large element is as difficult as solving the original problem), it seems to be perfectly suited for wave propagation where the element size is determined by the need to resolve the wave structure (control of the best approximation error).

Performance vs. traditional methods

While extremely stable in contrast to traditional Bubnov-Galerkin approximations for the Helmholtz problem, DPG forces us to consider the mixed problem in which we compute pressure, velocity, and additional fluxes. For a discretization of n elements of order p per wavelength, for a domain of m wavelengths, the DPG formulation requires $2(p+1)mn + 2mn$ degrees of freedom, while a standard H^1 conforming formulation involving only pressure requires pmn. After performing static condensation on the interior degrees of freedom, we are left with a system of dimension $2mn$ for DPG versus mn for the H^1 conforming method. This implies that the DPG method will be competitive when compared with standard finite elements only for large wavenumbers. Here we also remark that while static condensation for the H^1 conforming method requires us to avoid element sizes $h \approx \frac{j\pi}{k}, j \in \mathbb{N}$, in order to avoid the associated interior modes at the given wavenumber, in the mixed formulation which involves both p and u, there is no such trouble – the problem

$$\begin{aligned} ikp + u' &= 0, \\ iku + p' &= 0, \\ u(x_{i-1}) = u(x_i) = p(x_{i-1}) &= p(x_i) = 0 \end{aligned}$$

admits only the trivial solution. Hence, our element interior submatrices can always be factored without any concern of encountering interior modes, which should make static condensation, nested dissection, and other domain decomposition approaches robust.

Current and future work

The presented analysis and experiments are being extended to the two dimensional (2D) problem. Preliminary numerical results indicate that the presented methodology extends to multiple dimensions, with either zero, or numerically unobservable, phase error. A 2D version of the method on a structured rectangular mesh does not fit within the class of generalized finite element methods (GFEM) analyzed in [4]. Our current efforts concentrate on a 2D convergence analysis.

DPG methods can prove to be an attractive choice for high-frequency wave-propagation problems. Our methodology also offers the possibility of using plane waves or other waves (see, e.g. [15, 22]) for trial functions, with the hope of improving the approximation properties of the underlying space for a given problem. The theory of optimal test functions and optimal test norm continues to apply with better trial space choices. We emphasize nonetheless that the control of phase error is related to stability and *not approximability*, and the stability is controlled by the choice of test functions.

A. Proofs of the lemmas

Proof of Lemma 3.1. We need to prove that $\|q\|^2_{\tilde{V}} = \|ikq + q'\|^2_{L^2(0,1)} + \frac{1}{2}\|q\|^2_{L^2(0,1)}$ and $\|q\|^2_V = \|ikq + q'\|^2_{L^2(0,1)} + |q(1)|^2$ define equivalent norms on $H^1(0,1)$. Let $\tilde{q} = e^{ikx}q$. Since k is real,

$$\|\tilde{q}'\|^2_{L^2} = \|ikq + q'\|^2_{L^2}, \quad \|\tilde{q}\|^2_{L^2} = \|q\|^2_{L^2} \quad \text{and} \quad |\tilde{q}(1)|^2 = |q(1)|^2.$$

Thus, we only need to bound $\|\tilde{q}'\|_{L^2}^2 + |\tilde{q}(1)|^2$ above and below by $\|\tilde{q}'\|_{L^2}^2 + \frac{1}{2}\|\tilde{q}\|_{L^2}^2$. We use the Fundamental Theorem of Calculus $\tilde{q}(1) - \tilde{q}(x) = \int_x^1 \tilde{q}(s)\,ds$, together with standard techniques involving Young's inequality to estimate $|\tilde{q}(1)|^2$ and $\|\tilde{q}\|_{L^2}^2$. For every $\varepsilon > 0$ and $\delta > 0$,

$$\begin{aligned} |\tilde{q}(1)|^2 \leq \quad & (1+\varepsilon)\|\tilde{q}\|_{L^2}^2 + \tfrac{1+\varepsilon}{\varepsilon}\|\tilde{q}'\|_{L^2}^2, \\ \|\tilde{q}\|_{L^2}^2 \leq \quad & (1+\delta)|\tilde{q}(1)|^2 + \tfrac{1+\delta}{\delta}\|\tilde{q}'\|_{L^2}^2. \end{aligned}$$

Thus we already see that the norms $\|q\|_{\breve{V}}$ and $\|q\|_V$ are equivalent with constants of equivalence independent of k.

The rest of the proof is devoted to finding the constants stated in the lemma. To obtain the best constants, we first note that the above implies

$$\begin{aligned} \|q\|_V^2 &\leq \big(1 + \varepsilon^{-1}(1+\varepsilon)\big)\|\tilde{q}'\|_{L^2}^2 + (1+\varepsilon)\|\tilde{q}\|_{L^2}^2 && \text{(A.49)} \\ &\leq F_1(\varepsilon,\delta)\|\tilde{q}'\|_{L^2}^2 + F_2(\varepsilon,\delta)|\tilde{q}(1)|^2, && \text{(A.50)} \end{aligned}$$

where $F_1(\varepsilon,\delta) = \big(1 + \frac{1+\varepsilon}{\varepsilon} + (1+\varepsilon)\frac{1+\delta}{\delta}\big)$ and $F_2(\varepsilon,\delta) = (1+\varepsilon)(1+\delta)$. Then we minimize $F_2(\varepsilon,\delta)$ subject to the constraints $\varepsilon > 0, \delta > 0$, and $F_2(\varepsilon,\delta) = F_1(\varepsilon,\delta)$, to obtain $\varepsilon = \frac{\sqrt{2}}{2}$ and $\delta = 1+\sqrt{2}$. With these values, we have

$$\begin{aligned} \|q\|_V^2 &\leq \big(2+\sqrt{2}\big)\big(\|\tilde{q}'\|_{L^2}^2 + \frac{1}{2}\|\tilde{q}\|_{L^2}^2\big), && \text{by (A.49)}, \\ \big(2+\sqrt{2}\big)\big(\|\tilde{q}'\|_{L^2}^2 + \frac{1}{2}\|\tilde{q}\|_{L^2}^2\big) &\leq \big(2+\sqrt{2}\big)\big(\frac{2+\sqrt{2}}{2}\big)\big(\|\tilde{q}'\|_{L^2}^2 + |\tilde{q}(1)|^2\big), && \text{by (A.50)}. \end{aligned}$$

These two inequalities prove the lemma. □

Proof of Lemma 3.2. Let us prove that $\breve{V}_{hp} \subseteq V_{hp}$. Clearly this will imply that $\breve{\rho}_{hp} = \rho_{hp}$ due to unique solvability.

Let $q \in \breve{V}_{hp}$ denote the global optimal test function corresponding to $(\rho, \hat{\rho}) \in \breve{U}_{hp}$. It solves (3.32), i.e.,

$$\int_0^1 (ikq + q')\overline{(ik\delta q + \delta q')} + \frac{1}{2}\int_0^1 q\overline{\delta q} = \int_0^1 -\rho(\overline{ik\delta q + \delta q'}) + \hat{\rho}\overline{\delta q(1)},$$

for all $\delta q \in H^1(0,1)$. As ρ is smooth (a polynomial) within each element K_j, this variational equation translates into the following differential, boundary and interface equations:

$$\left\{\begin{array}{rll} -q'' - 2ikq' + (k^2 + \frac{1}{2})q &= \rho' + ik\rho & \text{in } (x_{j-1}, x_j) \text{ for } j = 1,\dots,n \\ [ikq + q' + \rho]_j &= 0 & \text{for } j = 1,\dots,n-1 \\ (ikq + q' + \rho)(1) &= \hat{\rho} & \\ -(ikq + q' + \rho)(0) &= 0. & \end{array}\right. \tag{A.51}$$

where, as before, the $[\,\cdot\,]_j$ denotes the jump of the argument at x_j.

Now, let $\delta q_j \in H^1(K_j)$. Multiplying $(\text{A.51})_1$ by δq_j, integrating over each element, and summing up over all elements, we get

$$\sum_{j=1}^n \left\{ -\int_{x_{j-1}}^{x_j} (ikq + q' + \rho)'\overline{\delta q_j} - \int_{x_{j-1}}^{x_j} \left(ik(q'+\rho) - (k^2 + \frac{1}{2})q\right)\overline{\delta q_j} \right\} = 0.$$

Integrating the first term by parts, using the continuity of $ikq+q'+\rho$ at element interfaces, and using the boundary conditions in $(\text{A.51})_{3,4}$, we obtain

$$\sum_{j=1}^{n}\left\{ \left(ikq+q', ik\delta q_j+\delta q_j'\right)_{L^2(K_j)} + \frac{1}{2}(q,\delta q_j)_{L^2(K_j)}\right\}$$
$$= \hat{\rho}\overline{\delta q_n}(1) - \sum_{j=1}^{n}\int_{x_{j-1}}^{x_j} \rho(\overline{ikq_j+q_j'}) + \sum_{j=1}^{n-1}(ikq+q'+\rho)(x_j)[\delta q]_j.$$

The left hand side equals $(q,\delta q)_{\tilde{V}}$, the inner product in (3.36) for the multielement case. The right hand side is the sesquilinear form of the DPG formulation (3.37) with

$$\begin{aligned}
&\rho_{hp} = \rho,\\
&\hat{\rho}_j^{hp} = (ikq+q'+\rho)(x_j), \quad \text{for } j=1,...,n-1,\\
&\hat{\rho}_n^{hp} = \hat{\rho}.
\end{aligned}$$

Therefore, the global optimal test function q is a linear combination of the local optimal test functions associated with ρ, $(ikq+q'+\rho)(x_j)$ and $\hat{\rho}$. □

Proof of Lemma 4.1. Let $r=\frac{v+q}{2}$ and $s=\frac{v-q}{2}$. Using the same idea as in the proof of Lemma 3.1, we set $\tilde{r}=e^{ikx}r$ and $\tilde{s}=e^{-ikx}s$, so

$$e^{-ikx}\tilde{r}' = ikr+r' \quad \text{and} \quad e^{ikx}\tilde{s}' = -iks+s'.$$

The norm $\|q,v\|_V^2$ can be expressed in terms of these new functions as

$$\begin{cases}
\|q,v\|_V^2 = & \|ik(r-s)+(r+s)'\|_{L^2}^2 + \|ik(r+s)+(r-s)'\|_{L^2}^2 + \left|2r(1)\right|^2 + |(r+s)(0)|^2\\
= & \left\|e^{-ikx}\tilde{r}'+e^{ikx}\tilde{s}'\right\|_{L^2}^2 + \left\|e^{-ikx}\tilde{r}'-e^{ikx}\tilde{s}'\right\|_{L^2}^2 + \left|2\tilde{r}(1)\right|^2 + |(\tilde{r}+\tilde{s})(0)|^2\\
= & 2\|\tilde{r}'\|_{L^2}^2 + 2\|\tilde{s}'\|_{L^2}^2 + \left|2\tilde{r}(1)\right|^2 + |(\tilde{r}+\tilde{s})(0)|^2.
\end{cases}$$

On one hand we have

$$\begin{aligned}
|2\tilde{r}(1)|^2 \le & \ (1+\varepsilon_1)\|2\tilde{r}\|_{L^2}^2 + 4\left(\tfrac{1+\varepsilon_1}{\varepsilon_1}\right)\|\tilde{r}'\|_{L^2}^2\\
|(\tilde{r}+\tilde{s})(0)|^2 \le & \ (1+\varepsilon_2)\|\tilde{r}+\tilde{s}\|_{L^2}^2 + \tfrac{1+\varepsilon_2}{\varepsilon_2}\|\tilde{r}'+\tilde{s}'\|_{L^2}^2\\
\le & \ (1+\varepsilon_2)\|\tilde{r}+\tilde{s}\|_{L^2}^2 + \tfrac{1+\varepsilon_2}{\varepsilon_2}\Big((1+\varepsilon_3)\|\tilde{r}'\|_{L^2}^2 + \tfrac{1+\varepsilon_3}{\varepsilon_3}\|\tilde{s}'\|_{L^2}^2\Big).
\end{aligned}$$

Observe that $\|2\tilde{r}\|_{L^2}^2 = \|v+q\|_{L^2}^2$ and $\|\tilde{r}+\tilde{s}\|_{L^2}^2 = \|v\cos(kx)+iq\sin(kx)\|_{L^2}^2$.

On the other hand, we have

$$\begin{aligned}
\|2\tilde{r}\|_{L^2}^2 \le & \ (1+\delta_1)|2\tilde{r}(1)|^2 + 4\left(\tfrac{1+\delta_1}{\delta_1}\right)\|\tilde{r}'\|_{L^2}^2\\
\|\tilde{r}+\tilde{s}\|_{L^2}^2 \le & \ (1+\delta_2)|(\tilde{r}+\tilde{s})(0)|^2 + \tfrac{1+\delta_2}{\varepsilon_2}\|\tilde{r}'+\tilde{s}'\|_{L^2}^2\\
\le & \ (1+\delta_2)|(\tilde{r}+\tilde{s})(0)|^2 + \tfrac{1+\delta_2}{\delta_2}\Big((1+\delta_3)\|\tilde{r}'\|_{L^2}^2 + \tfrac{1+\delta_3}{\varepsilon_3}\|\tilde{s}'\|_{L^2}^2\Big).
\end{aligned}$$

We choose $\varepsilon_1 = \varepsilon_2$, $\delta_1 = \delta_2$ and $\varepsilon_3 = \delta_3 = \sqrt{5} - 2$. As a consequence,

$$\left\{ \begin{array}{rcl} \|q, v\|_V^2 \leq & (1 + f_1(\varepsilon_1))\Big(2\|\tilde{r}'\|_{L^2}^2 + 2\|\tilde{s}'\|_{L^2}^2\Big) + f_2(\varepsilon_1)\Big(\|2\tilde{r}\|_{L^2}^2 + \|\tilde{r} + \tilde{s}\|_{L^2}^2\Big) \\ \\ \leq & (1 + f_1(\varepsilon_1) + f_2(\varepsilon_1) f_1(\delta_1))\Big(2\|\tilde{r}'\|_{L^2}^2 + 2\|\tilde{s}'\|_{L^2}^2\Big) \\ \\ & + f_2(\varepsilon_1) f_2(\delta_1)\Big(|2\tilde{r}(1)|^2 + |(\tilde{r} + \tilde{s})(0)|^2\Big), \end{array} \right. \tag{A.52}$$

where the functions of f_1 and f_2 are defined by :

$$f_1(x) = C\left(\frac{1+x}{x}\right), \quad f_2(x) = (1+x) \quad \text{and} \quad C = \frac{\sqrt{5}+3}{2}.$$

We arrive at the minimization problem :

$$\left\{ \begin{array}{rl} \min f_2(\varepsilon_1) f_2(\delta_1) & \\ \varepsilon_1 > & 0 \\ \delta_1 > & 0 \\ 1 + f_1(\varepsilon_1) + f_2(\varepsilon_1) f_1(\delta_1) - f_2(\varepsilon_1) f_2(\delta_1) = & 0, \end{array} \right.$$

whose solutions are $\varepsilon_1 = \sqrt{\frac{C}{C+1}}$ and $\delta_1 = f_1(\varepsilon_1)$. Replacing these values on the inequality (A.52) we obtain :

$$\left\{ \begin{array}{rcl} \|q, v\|_V^2 \leq & \sqrt{C+1}\big(\sqrt{C} + \sqrt{C+1}\big)\Big(2\|\tilde{r}'\|_{L^2}^2 + 2\|\tilde{s}'\|_{L^2}^2 + \frac{1}{C+1}\big(\|2\tilde{r}\|_{L^2}^2 + \|\tilde{r} + \tilde{s}\|_{L^2}^2\big)\Big) \\ \\ \leq & \sqrt{C+1}\big(\sqrt{C} + \sqrt{C+1}\big)\Big(1 + \sqrt{\frac{C}{C+1}}\Big)\|q, v\|_V^2. \end{array} \right.$$

Hence,

$$C_1 = \Big(1 + \sqrt{\tfrac{C}{C+1}}\Big)^{-1} = \sqrt{C+1}\big(\sqrt{C+1} - \sqrt{C}\big) = \tfrac{5+\sqrt{5}}{2} - (5 + 2\sqrt{5})^{\frac{1}{2}},$$

$$C_2 = \sqrt{C+1}\big(\sqrt{C+1} + \sqrt{C}\big) = \tfrac{5+\sqrt{5}}{2} + (5 + 2\sqrt{5})^{\frac{1}{2}}.$$

□

Acknowledgements

J. Zitelli was supported by an ONR Graduate Traineeship and CAM Fellowhip. I. Muga was supported by Sistema Bicentenario BECAS CHILE (Chilean Government). L. Demkowicz was supported by a Collaborative Research Grant from King Abdullah University of Science and Technology (KAUST). J. Gopalakrishnan was supported by the National Science Foundation under grant DMS-1014817.

References

[1] M. Ainsworth and H. Wajid. Optimally blended spectral-finite element scheme for wave propagation, and non-standard reduced integration. *University of Strathclyde Mathematics Research Report*, 12, 2009.

[2] I. Babuška, F. Ihlenburg, E.T. Paik, and S.A. Sauter. A generalized finite element method for solving the Helmholtz equation in two dimensions with minimal pollution. *Comput. Methods Appl. Mech. Engrg.*, 128:325–359, 1995.

[3] I. Babuska and J. M. Melenk. The partition of unity method. *International Journal of Numerical Methods in Engineering*, 40:727–758, 1996.

[4] Ivo M. Babuska and Stefan A. Sauter. Is the pollution effect of the FEM avoidable for the helmholtz equation considering high wave numbers? *SIAM J. Numer. Anal.*, 34(6):2392–2423, 1997.

[5] P. E. Barbone and I. Harari. Nearly H^1-optimal finite element methods. *Comput. Methods Appl. Mech. Engrg.*, 190:5679 – 5690, 2001.

[6] C.L. Bottasso, S. Micheletti, and R. Sacco. The discontinuous Petrov-Galerkin method for elliptic problems. *Comput. Methods Appl. Mech. Engrg.*, 191:3391–3409, 2002.

[7] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick. First-order system least squares for second-order partial differential equations. *SIAM J. Numer. Anal.*, 31:1785 – 1799, 1994.

[8] R. Courant and K. O. Friedrichs. *Supersonic Flow and Shock Waves*. Interscience Publishers, Inc., New York, N.Y., 1948.

[9] L. Demkowicz. Babuška $\Leftrightarrow$ Brezzi ? Technical report, ICES, 2006.

[10] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part I: The transport equation. Computer Methods in Applied Mechanics and Engineering, 199 (2010), pp. 1558—1572.

[11] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part II: Optimal test functions. Technical Report 16, ICES, 2009. In print (*Numer. Methods Partial Differential Equations*).

[12] L. Demkowicz, J. Gopalakrishnan, and A. Niemi. A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity. Technical Report 1, ICES, 2010. In review.

[13] L. Demkowicz and J. T. Oden. An adaptive characteristic Petrov-Galerkin finite element method for convection-dominated linear and nonlinear parabolic problems in one space variable. *Journal of Computational Physics*, 68(1):188–273, 1986.

[14] L. Demkowicz and J. T. Oden. An adaptive characteristic Petrov-Galerkin finite element method for convection-dominated linear and nonlinear parabolic problems in two space variables. *Comput. Methods Appl. Mech. Engrg.*, 55(1-2):65–87, 1986.

[15] C. Farhat, I. Harari, and L. P. Franca. The discontinuous enrichment method. *Computer Methods in Applied Mechanics and Engineering*, 190(48):6455 – 6479, 2001.

[16] Charbel Farhat, Isaac Harari, and Ulrich Hetmaniuk. A discontinuous galerkin method with lagrange multipliers for the solution of helmholtz problems in the mid-frequency regime. *Computer Methods in Applied Mechanics and Engineering*, 192(11-12):1389 – 1419, 2003.

[17] X. Feng and H. Wu. Discontinuous Galerkin methods for the Helmholtz equation with large wave number. *SIAM J. Numer. Anal.*, 47:2872 – 2896, 2009.

[18] Leopoldo P. Franca, Charbel Farhat, Antonini P. Macedo, and Michel Lesoinne. Residual-free bubbles for the Helmholtz equation. *Internat. J. Numer. Methods Engrg.*, 40(21):4003–4009, 1997.

[19] Dan Givoli. Nonlocal and semilocal optimal weighting functions for symmetric problems involving a small parameter. *Internat. J. Numer. Methods Engrg.*, 26(6):1281–1298, 1988.

[20] Isaac Harari. A survey of finite element methods for time-harmonic acoustics. *Computer Methods in Applied Mechanics and Engineering*, 195(13-16):1594 – 1607, 2006. A Tribute to Thomas J.R. Hughes on the Occasion of his 60th Birthday.

[21] Isaac Harari and Thomas J. R. Hughes. Finite element methods for the helmholtz equation in an exterior domain: Model problems. *Computer Methods in Applied Mechanics and Engineering*, 87(1):59 – 96, 1991.

[22] R. Hiptmair, A. Moiola, and I. Perugia. Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the p-version. Technical Report 20, Seminar for Applied Mathematics, ETH Zurich, 2009.

[23] T.J.R. Hughes and A. Brooks. A multidimensional upwind scheme with no crosswind diffusion. In *Finite Element Methods for Convection Dominated Flows (Papers, Winter Ann. Meeting Amer. Soc. Mech. Engrs., New York, 1979)*, volume 34 of *AMD*, pages 19–35, New York, 1979. Amer. Soc. Mech. Engrs. (ASME).

[24] F. Ihlenburg. *Finite Element Analysis of Acoustic Scattering*, volume 132 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1998.

[25] A. F. D. Loula and D. T. Fernandes. A quasi optimal Petrov-Galerkin method for Helmholtz problem. *Internat. J. Numer. Methods Engrg.*, 80:1595 – 1622, 2009.

[26] Assad A. Oberai and Peter M. Pinsky. A numerical comparison of finite element methods for the Helmholtz equation. *Journal of Computational Acoustics*, 8(1):211, 2000.

[27] J.T. Oden and L.F. Demkowicz. *Applied Functional Analysis for Science and Engineering*. Chapman & Hall/CRC Press, Boca Raton, 2010. Second edition.

[28] Lonny L. Thompson. A review of finite-element methods for time-harmonic acoustics. *The Journal of the Acoustical Society of America*, 119(3):1315–1330, 2006.

[29] Lonny L. Thompson and Peter M. Pinsky. A Galerkin least squares finite element method for the two-dimensional Helmholtz equation. *International Journal for Numerical Methods in Engineering*, 38:371 – 397, 1995.

[30] K. Yosida. *Functional Analysis*. Springer-Verlag, Berlin, 1995. Reprint of the sixth (1980) edition.