Dispersive and Dissipative Errors in the DPG Method with Scaled Norms for Helmholtz Equation

Jay Gopalakrishnan  
*Portland State University, gjay@pdx.edu*

Ignacio Muga  
*Pontificia Universidad Catolica de Chile*

Nicole Olivares  
*Portland State University*

Let us know how access to this document benefits you.

Follow this and additional works at: [http://pdxscholar.library.pdx.edu/mth_fac](http://pdxscholar.library.pdx.edu/mth_fac)

Part of the [Mathematics Commons](http://pdxscholar.library.pdx.edu/mth_fac)

Citation Details

DISPERSIVE AND DISSIPATIVE ERRORS IN THE DPG METHOD WITH SCALED NORMS FOR HELMHOLTZ EQUATION*

J. GOPALAKRISHNAN†, I. MUGA‡, AND N. OLIVARES†

This paper is dedicated to Leszek Demkowicz on the occasion of his 60th birthday

Abstract. This paper studies the discontinuous Petrov–Galerkin (DPG) method, where the test space is normed by a modified graph norm. The modification scales one of the terms in the graph norm by an arbitrary positive scaling parameter. The main finding is that as the parameter approaches zero, better results are obtained, under some circumstances, when the method is applied to the Helmholtz equation. The main tool used is a dispersion analysis on the multiple interacting stencils that form the DPG method. The analysis shows that the discrete wavenumbers of the method are complex, explaining the numerically observed artificial dissipation in the computed wave approximations. Since the DPG method is a nonstandard least-squares Galerkin method, its performance is compared with a standard least-squares method having a similar stencil.

Key words. least squares, dispersion, dissipation, quasi optimality, resonance, stencil

AMS subject classifications. 65N30, 35J05

DOI. 10.1137/130918186

1. Introduction. Discontinuous Petrov–Galerkin (DPG) methods were introduced in [8, 9, 10]. The DPG methods minimize a residual norm, so they belong to the class of least-squares Galerkin methods [3, 7, 14], although the functional setting in DPG methods is nonstandard. In this paper, we study a variant of the DPG method, referred to throughout as the DPG$_\varepsilon$ method, obtained by introducing an arbitrary parameter $\varepsilon > 0$ into the definition of the norm in which the residual is minimized. We study the properties of the resulting family of DPG methods when applied to the Helmholtz equation.

The DPG framework has already been applied to the Helmholtz equation in [12]. An error analysis with optimal error estimates was presented there. There are two major differences between the content of this paper and [12]. The first is the introduction of the abovementioned parameter, $\varepsilon$. When $\varepsilon = 1$, the method here reduces to that in [12]. The use of such scaling parameters was already advocated in [11] based on numerical experience. In this paper, we shall provide a theoretical basis for its use. The second difference with [12] is that in this contribution we perform a dispersion analysis of the DPG$_\varepsilon$ method, which uncovers several important properties of the method as $\varepsilon$ is varied.

Least-squares Galerkin methods are popular methods in scientific computation [3, 17]. They yield Hermitian positive definite systems, notwithstanding the indefiniteness of the underlying problem. Hence they are attractive from the point of view of

---

*Submitted to the journal’s Methods and Algorithms for Scientific Computing section April 23, 2013; accepted for publication (in revised form) September 25, 2013; published electronically January 2, 2014. This work was partially supported by the NSF under grant DMS-1318916, by the AFOSR under grant FA9550-12-1-0484, by the FONDECYT project 1130776, and by the CONICYT project Anillo ACT1118 (ANANUM).

†Portland State University, P.O. Box 751, Portland, OR 97207-0751 (gjay@pdx.edu, nmo@pdx.edu).

‡Pontificia Universidad Católica de Valparaíso, Casilla 4059, Valparaíso, Chile (ignacio.muga@ucv.cl)
solver design and many works have focused on this subject [18, 19]. However, as we shall see, for wave propagation problems, they yield solutions with heavy artificial dissipation. Since the DPG method is of the least-squares type, it also suffers from this problem. One of the goals of this paper is to show that by means of the \( \varepsilon \)-scaling, we can rectify this problem to some extent.

To explain this issue further, let us fix the specific boundary value problem we shall consider. Let \( A : H(\text{div}, \Omega) \times H^1(\Omega) \rightarrow L^2(\Omega)^N \times L^2(\Omega) \) denote the Helmholtz wave operator defined by

\[
A(\vec{v}, \eta) = (i\omega \vec{v} + \vec{\nabla} \eta, i\omega \eta + \vec{\nabla} \cdot \vec{v}).
\]

Here \( i \) denotes the imaginary unit, \( \omega \) is the wavenumber, and \( \Omega \) is a bounded open connected domain with Lipschitz boundary. All function spaces in this paper are over the complex field \( \mathbb{C} \). The Helmholtz equation takes the form \( A(\vec{u}, \phi) = f \) for some \( f \in L^2(\Omega)^N \times L^2(\Omega) \). Although we consider a general \( f \) in this paper, in typical applications, \( f = (\vec{0}, f) \) with \( f \in L^2(\Omega) \), in which case, eliminating the vector component \( \vec{u} \), we recover the usual second order form of the Helmholtz equation,

\[
-\Delta \phi - \omega^2 \phi = i\omega f \quad \text{on } \Omega.
\]

This must be supplemented with boundary conditions. The DPG method for the case of the impedance boundary conditions \( i\omega \phi + \partial \phi / \partial n = 0 \) on \( \partial \Omega \) was discussed in [12], but other boundary conditions are equally well admissible. In the present work, we consider the Dirichlet boundary condition

\[
\phi = 0 \quad \text{on } \partial \Omega.
\]

To deal with this boundary condition, we will need the space

\[
R = H(\text{div}, \Omega) \times H^1_0(\Omega).
\]

Thus, our boundary value problem reads as follows:

\[
\text{(1.4)} \quad \text{Find } (\vec{u}, \phi) \in R \text{ satisfying } A(\vec{u}, \phi) = f.
\]

It is well known [16] that except for \( \omega \) in \( \Sigma \), an isolated countable set of real values, this problem has a unique solution. We assume henceforth that \( \omega \) is not in \( \Sigma \).

Before studying the DPG \( \varepsilon \) method for (1.4), it is instructive to examine the simpler \( L^2 \) least-squares Galerkin method. Set \( R_h \subset R \) to the Cartesian product of the lowest order Raviart–Thomas and Lagrange spaces, together with the boundary condition in \( R \). The method finds \( (\vec{u}_h^{ls}, \phi_h^{ls}) \in R_h \) such that

\[
\text{(1.5)} \quad (\vec{u}_h^{ls}, \phi_h^{ls}) = \arg \min_{w \in R_h} \| f - Aw \|.
\]

Throughout, \( \| \cdot \| \) denotes the \( L^2(\Omega) \)-norm, or the natural norm in the Cartesian product of several \( L^2(\Omega) \) component spaces. The method (1.5) belongs to the so-called FOSLS [7] class of methods.

Although (1.5) appears at first sight to be a reasonable method, computations yield solutions with artificial dissipation. For example, suppose we use (1.5), appropriately modified to include nonhomogeneous boundary conditions, to approximate a plane wave propagating at angle \( \theta = \pi/8 \) in the unit square. A comparison between the real parts of the exact solution (in Figure 1.1(a)) and the computed solution (in Figure 1.1(b)) shows that the computed solution dissipates at interior mesh points.
A22

J. GOPALAKRISHNAN, I. MUGA, AND N. OLIVARES

Fig. 1.1. Approximations to a plane wave computed using a uniform mesh of square elements of size $h = 1/48$ (about sixteen elements per wavelength). Artificial dissipation is visible in Figures 1.1(b) and 1.1(c). (The parameter $r$ in the DPG$_\varepsilon$ method, defined in section 2.4, is set to 3 for these computations.)

The same behavior is observed for the lowest order DPG$_\varepsilon$ method (which has $\varepsilon = 1$) in Figure 1.1(c). The DPG$_\varepsilon$ method with $\varepsilon = 10^{-6}$, however, gave a solution (in Figure 1.1(d)) that is visually indistinguishable from the exact solution. (Note that, for the DPG$_\varepsilon$ method with $\varepsilon = 1$, the numerical results presented in [12] show much better performance, because slightly higher order spaces were used there. Instead, in this paper, we have chosen to study the DPG$_\varepsilon$ method with the lowest possible order of approximation spaces to reveal the essential difficulties with minimal computational effort.)

The situation in Figures 1.1(b) and 1.1(c) improves when more elements per wavelength are used. This is not surprising in view of the asymptotic error estimates of the methods. To give an example of such an error estimate, consider the case of the impedance boundary conditions considered in [12]. It is proven there that there is a constant $C > 0$, independent of $\omega$ and mesh size $h$, such that the lowest order
DISPERSION ANALYSIS FOR DPG METHODS

DPG solution \((\vec{u}_h, \phi_h)\) satisfies

\[
\|\vec{u} - \vec{u}_h\| + \|\phi - \phi_h\| \leq C\omega^2 h
\]

for a plane wave solution. A critical ingredient in this analysis is the estimate

\[
\|w\| \leq C'\|Aw\|,
\]

which, as shown in [12, Lemmas 4.2 and 4.3], holds for all \(w\) in the analogue of \(R\) with impedance boundary conditions. Although the analysis in [12] was for the impedance boundary condition, similar techniques apply to the Dirichlet boundary condition as well, leading to (1.6). As more elements per wavelength are used, \(\omega h\) decreases, so (1.6) guarantees that the situation in Figure 1.1(c) will improve.

The analysis for the \(L^2\) least-squares method is even easier: By (1.5), \(\|f - A(\vec{u}_{ls}^h, \phi_{ls}^h)\| \leq \|A(\vec{u} - \vec{w}_h, \phi - \psi_h)\|\) for any \((\vec{w}_h, \psi_h) \in \mathcal{R}_h\). Hence, applying (1.7) to the error \(e = (\vec{u} - \vec{u}_{ls}^h, \phi - \phi_{ls}^h)\) and noting that the residual is \(Ae = f - A(\vec{u}_{ls}^h, \phi_{ls}^h)\), we obtain \(\|e\| \leq C'\|A(\vec{u} - \vec{w}_h, \phi - \psi_h)\|\). By standard approximation estimates, we then conclude that there is a \(C > 0\) independent of \(\omega\) and \(h\) such that

\[
\|\vec{u} - \vec{u}_{ls}^h\| + \|\phi - \phi_{ls}^h\| \leq C\omega^2 h.
\]

This simple technique of analysis of \(L^2\)-based least-squares methods is well known (see, e.g., [17, pp. 70–71]). As with (1.6), the estimate (1.8) implies that as the number of elements per wavelength is increased, \(\omega h\) decreases, and the situation in Figure 1.1(b) must improve.

Yet, Figures 1.1(b) and 1.1(c) show that the accuracy of these methods fails to be competitive with standard methods when the number of elements per wavelength is small. The figures also illustrate one of the difficulties with asymptotic error estimates like (1.7) and (1.8). Having little knowledge of the size of \(C\), we cannot predict the performance of the method on coarse meshes. Motivated by this difficulty, one of the theorems we present (Theorem 3.1) will give a better idea of the constant involved as \(\varepsilon \to 0\). Also note that the above indicated error analyses do not give us a quantitative measure of differences in wave speeds between the computed and exact waves. This motivates the dispersion analysis we present in this paper, which will address the issue of wave speed discrepancies.

We should note that there are alternative methods of the least-squares type that exhibit better performance than the standard \(L^2\)-based least-squares method. Some are based on adding further terms to the residual to be minimized (e.g., to control the curl of the vector equation [18]). Another avenue explored by others, and closer to the subject of this paper, is the idea of minimizing the residual in a dual norm [4, 5]. The main difference with our method is that our dual norms are locally computable in contrast to their nonlocal norms. This is achieved by using an ultraweak variational setting. The domain and codomain of the operator in the least-squares minimization associated with the DPG method are nonstandard, as we shall see next.

2. The DPG method for the Helmholtz equation. In this section, we briefly review the method for the Helmholtz equation introduced in [12]. We then show exactly where the parameter \(\varepsilon\) is introduced to get the variant of the method that we intend to study.

Let \(\Omega_h\) be a disjoint partitioning of \(\Omega\) into open elements \(K\) such that \(\Omega = \bigcup_{K \in \Omega_h} K\). The shape of the mesh elements in \(\Omega_h\) is unimportant for now, except that
we require their boundaries $\partial K$ to be Lipschitz so that traces make sense. Let
\begin{equation}
V = H(\text{div}, \Omega_h) \times H^1(\Omega_h),
\end{equation}
where
\begin{align*}
H(\text{div}, \Omega_h) &= \{ \vec{v} : \vec{v}|_K \in H(\text{div}, K), \forall K \in \Omega_h \}, \\
H^1(\Omega_h) &= \{ v : v|_K \in H^1(K), \forall K \in \Omega_h \}.
\end{align*}

Let $A_h : V \to L^2(\Omega)^N \times L^2(\Omega)$ be defined in the same way as $A$ in (1.1), except the derivatives are taken element by element, i.e., on each $K \in \Omega_h$, we have $A_h(\vec{v}, \eta)|_K = (i\omega \vec{v}|_K + \vec{\nabla} \eta|_K, i\omega \eta|_K + \vec{\nabla} \cdot \vec{v}|_K)$.

### 2.1. Integration by parts.

The following basic formula that we shall use is obtained simply by integrating by parts each of the derivatives involved:
\begin{equation}
\int_D A(\vec{w}, \psi) \cdot (\vec{v}, \eta) = -\int_D (\vec{w}, \psi) \cdot A(\vec{v}, \eta) + \int_{\partial D} (\vec{w} \cdot \vec{n}) \eta + \int_{\partial D} \psi (\vec{v} \cdot \vec{n})
\end{equation}

for smooth functions $(\vec{w}, \psi)$ and $(\vec{v}, \eta)$ and domains $D$ with Lipschitz boundary. Above, overlines denote complex conjugations and the integrals use the appropriate Lebesgue measure. Note that we use the notation $\vec{n}$ throughout to generically denote the outward unit normal on various domains—the specific domain will be clear from context—e.g., in (2.2), it is $D$. Introducing the following abbreviated notations for tuples $\vec{w} = (\vec{w}, \psi)$ and $\vec{v} = (\vec{v}, \eta)$,
\begin{align*}
\langle \vec{w}, \vec{v} \rangle_h &= \sum_{K \in \Omega_h} \int_K (\vec{w} \cdot \vec{v}) + \psi \eta, \\
\langle \vec{w}, \vec{v} \rangle_h &= \sum_{K \in \Omega_h} \int_{\partial K} (\vec{w} \cdot \vec{n}) \eta + \int_{\partial K} \psi (\vec{v} \cdot \vec{n}),
\end{align*}

we can rewrite (2.2), applied element by element, as
\begin{equation}
\langle A \vec{w}, \vec{v} \rangle_h = -\langle \vec{w}, A_h \vec{v} \rangle_h + \langle \vec{w}, \vec{v} \rangle_h.
\end{equation}

By density, (2.3) holds for all $\vec{w} \in H(\text{div}, \Omega) \times H^1(\Omega)$ and all $\vec{v} \in V$. Then, $\langle \cdot, \cdot \rangle_h$ must be interpreted using the appropriate duality pairing as the last term in (2.3) contains interelement traces on $\partial \Omega_h = \{ \partial K : K \in \Omega_h \}$.

It will be convenient to introduce notation for such traces. Define
\begin{equation}
\text{tr}_h : H(\text{div}, \Omega) \times H^1(\Omega) \to \prod_K H^{-1/2}(\partial K) \vec{n} \times H^{1/2}(\partial K)
\end{equation}
as follows. For any $(\vec{w}, \psi) \in H(\text{div}, \Omega) \times H^1(\Omega)$, the restriction of $\text{tr}_h(\vec{w}, \psi)$ on the boundary of any mesh element $\partial K$ takes the form
\begin{equation}
((\vec{w} \cdot \vec{n})\vec{n} |_{\partial K}, \psi |_{\partial K}) \in H^{-1/2}(\partial K) \vec{n} \times H^{1/2}(\partial K).
\end{equation}

Although the meaning of $H^{-1/2}(\partial K) \vec{n}$ is more or less self-evident, to include a proper definition, first let $Z$ denote the space of all functions of the form $\xi \vec{n}$, where $\xi$ is in $H^{1/2}(\partial K)$, normed by $\| \xi \vec{n} \|_Z = \| \xi \|_{H^{1/2}(\partial K)}$. Let $Z'$ denote the dual space of $Z$. Now, consider the map $M \vec{q} = (\vec{q} \cdot \vec{n})\vec{n} |_{\partial K}$, defined for smooth functions $\vec{q}$ on $K$. Since
\begin{equation}
\int_{\partial K} M \vec{q} \cdot \xi \vec{n} = \int_{\partial K} (\vec{q} \cdot \vec{n})\vec{n} |_{\partial K},
\end{equation}

2.2. An ultraweak formulation. The boundary value problem we wish to approximate is (2.5). Recall the definition of \( R \) in (1.3). To deal with the Dirichlet boundary condition, we will need the trace space

\[
Q = \text{tr}_h(R).
\]

To derive the DPG method for

\[
\begin{align*}
(2.5a) & \quad A(\vec{u}, \phi) = f \quad \text{on } \Omega, \\
(2.5b) & \quad \phi = 0 \quad \text{on } \partial \Omega,
\end{align*}
\]

we use the integration by parts by formula (2.3) to get

\[
- \langle (\vec{u}, \phi), A_h(\vec{v}, \eta) \rangle_h + \langle \text{tr}_h(\vec{u}, \phi), (\vec{v}, \eta) \rangle_h = \langle f, (\vec{v}, \eta) \rangle_h
\]

for all \((\vec{v}, \eta) \in V\). Now we let \( \text{tr}_h(\vec{u}, \phi) \) be an independent unknown \((\hat{u}, \hat{\phi})\) in \( Q \). Defining the bilinear form \( b((\vec{u}, \phi, \hat{u}, \hat{\phi}), (\vec{v}, \eta)) = - \langle (\vec{u}, \phi), A_h(\vec{v}, \eta) \rangle_h + \langle (\vec{u}, \hat{\phi}), (\vec{v}, \eta) \rangle_h \), we obtain the ultraweak formulation of [12]: Find \( u = (\vec{u}, \phi, \hat{u}, \hat{\phi}) \) in

\[
U = L^2(\Omega)^N \times L^2(\Omega) \times Q
\]

satisfying

\[
(2.6) \quad b(u, v) = \langle f, v \rangle_h \quad \forall v \in V.
\]

The wellposedness of this formulation was proved in [12] for the case of impedance boundary conditions. As is customary, we refer to the solution component \( \hat{u} \) as the numerical flux and \( \hat{\phi} \) as the numerical trace.

2.3. The DPG\( \varepsilon \) method. Let \( U_h \subset U \) be a finite dimensional trial space. The DPG method finds \( u_h \) in \( U_h \) satisfying

\[
(2.7) \quad b(u_h, v_h) = \langle f, v_h \rangle_h
\]

for all \( v_h \) in the test space \( V_h \), defined by

\[
(2.8) \quad V_h = TU_h,
\]

where \( T : U \to V \) is defined by

\[
(2.9) \quad \langle Tw, v \rangle_V = b(w, v) \quad \forall v \in V,
\]

and the \( V \) inner product \( \langle \cdot, \cdot \rangle_V \) is the inner product generated by the norm

\[
(2.10) \quad \|v\|^2_V = \|A_h v\|^2 + \varepsilon^2 \|v\|^2.
\]
Here \( \varepsilon > 0 \) is an arbitrary scaling parameter. Note that when \( \varepsilon = 1 \), (2.10) defines a graph norm on \( V \). The case \( \varepsilon = 1 \), analyzed in [12], is the standard DPG method. The general case, which we refer to as the DPG\(_{\varepsilon} \) method, will be analyzed in the next section.

It is easy to reformulate the DPG\(_{\varepsilon} \) method as a residual minimization problem. (All DPG methods with test spaces as in (2.9) minimize a residual as already pointed out in [10].) Letting \( V' \) denote the dual space of \( V \), normed with \( \| \cdot \|_{V'} \), we define \( F \in V' \) by \( F(v) = (f, v)_{\Omega} \). Then letting \( B : U \to V' \) denote the operator generated by the above defined \( b(\cdot, \cdot) \), i.e., \( Bw(v) = b(w, v) \) for all \( w \in U \) and \( v \in V' \), one can immediately see that \( u_h \) solves (2.7) if and only if

\[
u_h = \arg \min_{w \in U_h} \| Bw - F \|_{V'} \implies \quad v_h \in U_h.
\]

This norm highlights the difference between the DPG method and the previously discussed standard \( L^2 \)-based least-squares method (1.5).

### 2.4. Inexactly computed test spaces.

A basis for the test space \( V_h \), defined in (2.8), can be obtained by applying \( T \) to a basis of \( U_h \). One application of \( T \) requires solving (2.9), which although local (calculable element by element), is still an infinite dimensional problem. Accordingly a practical version of the DPG\(_{\varepsilon} \) method uses a finite dimensional subspace \( V^r \subset V \) and replaces \( T \) by \( T^r : U \to V^r \) defined by

\[
\langle T^r w, v \rangle_V = b(w, v) \quad \forall v \in V^r.
\]

In computations, we then use, in place of \( V_h \), the inexactly computed test space \( V^r_h \equiv T^r U_h \), i.e., the practical DPG method finds \( u^r_h \) in \( U_h \) satisfying

\[
(2.12) \quad b(u^r_h, v) = (f, v)_h \quad \forall v \in V^r_h.
\]

For the Helmholtz example, we set \( V^r \) as follows: Let \( Q_{l,m} \) denote the space of polynomials of degree at most \( l \) and \( m \) in \( x_1 \) and \( x_2 \), resp. Let \( RT_r \equiv Q_{r,r-1} \times Q_{r-1,r} \) denote the Raviart–Thomas subspace of \( H(\text{div}, K) \). We set

\[
V^r = \{ v : v|_K \in RT_r \times Q_{r,r} \}.
\]

Clearly, \( V^r \subset H(\text{div}, \Omega_h) \times H^1(\Omega_h) \). Later, we shall solve (2.12) using \( r \geq 2 \) and report the numerical results. It is easy to see using the Fortin operators developed in [15] that \( T^r \) is injective for \( r \geq 2 \), which implies that (2.12) yields a positive definite system. However, a complete analysis using [15] tracking \( \omega \) and \( r \) dependencies remains to be developed, and is not the subject of this paper.

### 3. Analysis of the DPG\(_{\varepsilon} \) method.

The purpose of this section is to study how the stability constant of the DPG\(_{\varepsilon} \) method (2.7) depends on \( \varepsilon \). The analysis in this section provides the theoretical motivation to introduce the scaling by \( \varepsilon \) into the DPG setting.

#### 3.1. Assumption

The analysis is under the already placed assumption that the boundary value problem (2.5) is uniquely solvable. We will now need a quantitative form of this assumption. Namely, there is a constant \( C(\omega) > 0 \), possibly depending on \( \omega \), such that the solution of (2.5) satisfies

\[
\| (\hat{u}, \phi) \| \leq C(\omega)\| f \|.
\]
One expects $C(\omega)$ to become large as $\omega$ approaches any of the resonances in $\Sigma$. For any $(\vec{r}, \psi) \in R$, choosing $f = A(\vec{r}, \psi)$ and applying the above inequality, we obtain

$$\|\langle \vec{r}, \psi \rangle\| \leq C(\omega)\|A(\vec{r}, \psi)\| \quad \forall (\vec{r}, \psi) \in R. \quad (3.1)$$

This is the form in which we will use the assumption.

Note that in the case of the impedance boundary condition, the unique solvability assumption can be easily verified [20] for all $\omega$. Furthermore, when that boundary condition is imposed, for instance, on the boundary of a convex domain, the estimate (3.1) is proved in [12, Lemmas 4.2 and 4.3] using a result of [20]. The resulting constant $C(\omega)$ is bounded independently of $\omega$. However, we cannot expect this independence to hold for the Dirichlet boundary condition (1.2) we are presently considering.

Finally, let us note that the ensuing analysis applies equally well to the impedance boundary condition: We only need to replace the space $R$ considered here by that in [12] and assume (3.1) for all functions in the revised $R$.

### 3.2. Quasi optimality.

It is well known that if there are positive constants $C_1$ and $C_2$ such that

$$C_1\|v\|_V \leq \sup_{w \in U} \frac{|b(w, v)|}{\|w\|_U} \leq C_2\|v\|_V \quad \forall v \in V, \quad (3.2)$$

then a quasi-optimal error estimate

$$\|u - u_h\|_U \leq \frac{C_2}{C_1} \inf_{w \in W_h} \|u - w\|_U \quad (3.3)$$

holds. This follows from [12, Theorem 2.1], or from the more general result of [15, Theorem 2.1], after noting that the following uniqueness condition holds: Any $w \in U$ satisfying $b(w, v) = 0$ for all $v \in V$ vanishes. (Since this uniqueness condition can be proved in [12, Lemma 4.1], we shall not dwell on it here.)

Accordingly, the remainder of this section is devoted to proving (3.2), tracking the dependence of constants with $\varepsilon$, and using the $U$-norm we define below. First, let

$$\|\langle \vec{r}, \psi \rangle\|_R = \frac{1}{\varepsilon}\|A(\vec{r}, \psi)\|_R. \quad (3.4)$$

By virtue of (3.1), this is clearly a norm under which the space $R$, defined in (1.3), is complete. The space $Q$ in (2.4) is normed by the quotient norm, i.e., for any $\hat{q} \in Q$,

$$\|\hat{q}\|_Q = \inf \{\|r\|_R : \forall r \in R \text{ such that } \text{tr}_h r = \hat{q}\}.$$ 

The function in $R$ which achieves the infimum above defines an extension operator $E : Q \to R$ that is a continuous right inverse of $\text{tr}_h$ and satisfies

$$\|E\hat{q}\|_R = \|\hat{q}\|_Q. \quad (3.4)$$

With these notations, we can now define the norm on the trial space by

$$\|(w, \psi, \hat{w}, \hat{\psi})\|_U^2 = \|(w, \psi)\|_U^2 + \|(\hat{w}, \hat{\psi})\|_Q^2. \quad (3.4)$$

The following theorem is proved by extending the ideas in [12] to the DPG$_{\varepsilon}$ method.

**Theorem 3.1.** Suppose (3.1) holds and $c = C(\omega)(C(\omega)\varepsilon/2 + \sqrt{1 + C(\omega)^2\varepsilon^2/4})$. Then the inf-sup condition in (3.2) holds with $C_1 = 1/\sqrt{1 + c\varepsilon}$ and the continuity...
condition in (3.2) holds with $C_2 = \sqrt{1 + c\varepsilon}$. Hence, the DPG solution admits the error estimate

$$\|u - u_h\|_U \leq (1 + c\varepsilon) \inf_{w \in U} \|u - w\|_U.$$  

Proof. We first prove the continuity estimate. Let $(w, \tilde{q}) \in U$ and let $v \in V$. We use the abbreviated notation $\hat{q} = (\hat{w}, \hat{\psi}), w = (w, \psi)$, and $v = (\hat{v}, \eta)$. By (3.1) and (3.4),

$$(3.5) \quad \|E\hat{q}\| \leq C(\varepsilon)\varepsilon\|\hat{q}\|_Q, \quad \|A\hat{E}\hat{q}\| = \varepsilon\|\hat{q}\|_Q.$$  

The extension $E$ can be used to rewrite $b((w, \tilde{q}), v) = -\langle w, A_h v \rangle_h + \langle E\hat{q}, A_h v \rangle_h + \langle A\hat{E}\hat{q}, v \rangle_h$. Then, applying the Cauchy–Schwarz inequality, and using (3.5), we have

$$(3.6) \quad |b((w, \tilde{q}), v)| \leq \|w\|_U\|A_h v\| + C(\varepsilon)\varepsilon\|\hat{q}\|_Q\|A_h v\| + \varepsilon\|\hat{q}\|_Q\|v\|$$

where $t^2 = \|A_h v\|^2 + (C(\varepsilon)\varepsilon\|A_h v\| + \varepsilon\|v\|)^2$. With $a = C(\varepsilon)\varepsilon\|A_h v\|$ and $b = \varepsilon\|v\|$ we apply the inequality $(a + b)^2 \leq (1 + \alpha^2)a^2 + (1 + \alpha^{-2})b^2$ to obtain

$$(3.7) \quad (1 + \alpha^2)C(\varepsilon)^2\varepsilon^2 = \alpha^{-2} = c\varepsilon$$

for any $\alpha > 0$. Set $\alpha^2 = -1/2 + \sqrt{1/4 + C(\varepsilon)^{-2}\varepsilon^{-2}}$, so that

$$t^2 \leq (1 + (1 + \alpha^2)C(\varepsilon)^2\varepsilon^2)\|A_h v\|^2 + (1 + \alpha^{-2})\varepsilon^2\|v\|^2$$

with $c$ as in the statement of the theorem. Hence, $t^2 \leq (1 + c\varepsilon)^2\|v\|^2$. Returning to (3.6),

$$|b((w, \tilde{q}), v)| \leq C_2\|b((w, \tilde{q}), v)\|_U\|v\|$$

with $C_2 = \sqrt{1 + c\varepsilon}$. This verifies the upper inequality of (3.2).

To prove the lower inequality of (3.2), let $r$ be the unique function in $R$ satisfying $Ar = v$ for any given $v \in V$. Then, by (3.1),

$$(3.8) \quad \|r\| \leq C(\varepsilon)\|v\|.$$  

Also, since $\|Ar\| = \|v\|$, letting $\hat{r} = \text{tr}_h r$, we have

$$(3.9) \quad \|\hat{r}\|_Q = \frac{1}{\varepsilon}\|A\hat{E}\hat{r}\| \leq \frac{1}{\varepsilon}\|Ar\| = \frac{1}{\varepsilon}\|v\|.$$  

By (2.3), we have $(Ar, v)_h = -\langle r, A_h v \rangle_h + \langle \hat{r}, v \rangle_h$, so

$$(3.10) \quad \|v\|^2_{\hat{v}} = \varepsilon^2\|v\|^2 + \|A_h v\|^2 = \varepsilon^2 b((z, \hat{r}), v),$$

where $z = r - \varepsilon^{-2}A_h v$, a function that can be bounded using (3.8), as follows:

$$\|z\|^2 \leq (1 + \alpha^2)\|r\|^2 + (1 + \alpha^{-2})\varepsilon^{-4}\|A_h v\|^2$$

$$\leq (1 + \alpha^2)C(\varepsilon)^2\|v\|^2 + (1 + \alpha^{-2})\varepsilon^{-4}\|A_h v\|^2$$
for any $\alpha > 0$. Choosing $\alpha$ as in (3.7) and using (3.8)–(3.9),

$$
\varepsilon^4 \| (z, \hat{r}) \|^2_U = \varepsilon^4 \| z \|^2 + \varepsilon^4 \| \hat{r} \|^2_Q \\
\leq (1 + (1 + \alpha^2) C(\omega)^2 \varepsilon^2) \varepsilon^2 \| v \|^2 + (1 + \alpha^{-2}) \| A_h v \|^2 \\
\leq (1 + c \varepsilon) \left( \varepsilon^2 \| v \|^2 + \| A_h v \|^2 \right).
$$

(3.11)

Returning to (3.10), we now have

$$
\| v \|^2_U = \frac{b((z, \hat{r}), v)}{\| (z, \hat{r}) \|^2_U} \varepsilon^2 \| (z, \hat{r}) \|^2_U \leq \left( \sup_{x \in U} \frac{|b(x, v)|}{\| x \|^2_U} \right) \sqrt{1 + c \varepsilon} \| v \|_V
$$

by virtue of (3.11), verifying the lower inequality of (3.2) with $C_1 = 1/\sqrt{1 + c \varepsilon}$. \hfill \Box

**Remark 3.2.** Although we presented the above result only for the Helmholtz equation, the ideas apply more generally. It seems possible to prove a similar result abstractly, e.g., using the abstract setting in [6], for any DPG application that uses a scaled graph norm analogous to (2.10) (with the wave operator $A_h$ replaced by suitable others).

### 3.3. Discussion

Theorem 3.1 shows that the use of the $\varepsilon$-scaling in the test norm can ameliorate some stability problems, e.g., those that can arise from large $C(\omega)$.

Observe that the best possible value for the constant $C_2/C_1$ in (3.3) is 1. Indeed, if $C_2/C_1$ equals 1, then the computed solution $u_h$ coincides with the best approximation to $u$ from $U_h$. Theorem 3.1 shows that the quasi-optimality constant of the DPG method approaches the ideal value of 1 as $\varepsilon \to 0$.

However, since the norms depend on $\varepsilon$, we must further examine the components of the error separately, by defining

$$
\varepsilon^2 = \| \bar{u} - \bar{u}_h \|^2 + \| \phi - \phi_h \|^2, \\
\hat{\varepsilon}^2 = \| AE(\hat{u} - \hat{u}_h, \hat{\phi} - \hat{\phi}_h) \|^2.
$$

(3.12a)

(3.12b)

The estimate of Theorem 3.1 implies that

$$
\varepsilon^2 + \hat{\varepsilon}^2 \leq (1 + c \varepsilon)^2 \left( a^2 + \hat{a}^2 \right),
$$

where $a$ and $\hat{a}$ are the best approximation errors defined by

$$
a^2 = \inf_{(\bar{u}, \bar{\phi}), \bar{\psi} \in U_h} \| \bar{u} - \bar{u} \|^2 + \| \phi - \psi \|^2, \\
\hat{a}^2 = \inf_{(0, 0, \bar{w}, \bar{\psi}) \in U_h} \| AE(\bar{u} - \bar{w}, \hat{\phi} - \hat{\psi}) \|^2.
$$

(3.14)

Note that $E$ is independent of $\varepsilon$.

We want to compare the error bounds for the numerical fluxes and traces in the $\varepsilon = 1$ case with the case of $0 < \varepsilon \ll 1$. To distinguish these cases we will denote the error defined in (3.12b) by $\hat{e}_1$ when $\varepsilon = 1$. Clearly, (3.13) implies

$$
\hat{e}_1^2 \leq (1 + c \varepsilon)^2 \left( a^2 + \hat{a}^2 \right).
$$

(3.15)

For the other case, (3.13) implies, after multiplying through by $\varepsilon^2$,

$$
\varepsilon^2 \leq (1 + c \varepsilon)^2 \left( \varepsilon^2 a^2 + \hat{a}^2 \right).
$$
Comparing this with (3.15), and noting that $a$ and $\hat{a}$ remain the same for different $\varepsilon$, we find that the DPG$_{\varepsilon}$ errors for fluxes and traces admit a better bound for smaller $\varepsilon$ in an $\varepsilon$-independent norm. Whether the actually observed numerical error improves will be investigated through the dispersion analysis presented in a later section, as well as in the next subsection.

### 3.4. Numerical illustration

Theorem 3.1 partially explains a numerical observation we now report. We implemented the DPG$_{\varepsilon}$ method by setting the parameter $r = 3$ (see section 2.4) and computed $u^e_h = (\overrightarrow{u}^e_h, \phi^e_h, \hat{u}^e_h, \hat{\phi}^e_h)$. In analogy with (3.12), define the discretization errors $e_r$ and $\hat{e}_r$ by $e_r^2 = \|\overrightarrow{u} - \overrightarrow{u}^e_h\|^2 + \|\phi - \phi^e_h\|^2$ and $\hat{e}_r^2 = \|AE(\hat{u} - \hat{u}^e_h, \phi - \hat{\phi}^e_h)\|^2$. Although Theorem 3.1 suggests an investigation of

$$\frac{\|u - u^e_h\|_U}{\inf_{w \in U_h} \|u - w\|_U} = \left(\frac{e_r^2 + (\hat{e}_r/\varepsilon)^2}{a^2 + (\hat{a}/\varepsilon)^2}\right)^{1/2},$$

due to the difficulty of applying the extension operator $E$ in practice, we have investigated the ratio $e_r/a$ as a function of $\omega$. Recall that $a$ is the $L^2(\Omega)$ best approximation error defined in (3.14), so $e_r/a$ measures how close the discretization errors are to the best possible.

For a range of wavenumbers $\omega$, we chose the data $f = (\overrightarrow{0}, f)$ so that the exact solution to (2.5) on the unit square would be $(\overrightarrow{u}, \phi) = (\frac{1}{\omega} \nabla \phi, \phi)$, with $\phi = x(1-x)y(1-y)$. Each resulting boundary value problem was then solved using the DPG$_{\varepsilon}$ method with $\varepsilon = 10^{-n}, n = 0, 1, \ldots, 4$, on a fixed mesh of $h = 1/16$ and the corresponding discretization errors $e_r$ were collected.

The resulting ratios $e_r/a$ are plotted as a function of $\omega$ in Figure 3.1 for a few $\varepsilon$ values. First of all, observe that the graph of the ratio begins close to the optimal value of one for all $\varepsilon$ values in the figure. Next, observe that the ratio spikes up as $\omega$ approaches the exact resonance value $\omega = \pi \sqrt{2} \approx 4.44$, where $C(\omega)$ is infinity. It is interesting to look at the points near (but not at) the resonance. Observe that as $\varepsilon$ is decreased, the DPG$_{\varepsilon}$ method exhibits a “regularizing” effect at points near the resonance, e.g., at $\omega = 5$, the values of $e_r/a$ are closer to 1 for smaller $\varepsilon$. It therefore seems advantageous to use smaller $\varepsilon$ for problems near resonance.

![Figure 3.1](image-url)
The theoretical explanation for this numerical observation would be complete (by virtue of Theorem 3.1), if we had computed using the exact DPG test spaces \((r = \infty)\), instead of the inexactely computed spaces \((r = 3)\). Certain discrete effects arising due to this inexact computation of test spaces will be presented in a later section.

4. Lowest order stencil. Consider the example of square two dimensional elements. The lowest order case of the DPG method is obtained using \(Q(\partial K) = \{(\hat{\omega}, \hat{\psi}) : \hat{\omega} \text{ is constant on each edge of } \partial K, \hat{\psi} \text{ is linear on each edge of } \partial K, \text{ and } \hat{\psi} \text{ is continuous on } \partial K\} \). Let \(S(K) = \{(\bar{\omega}, \bar{\psi}) : \bar{\omega} \text{ and } \bar{\psi} \text{ are constant (vector and scalar, resp.) functions on } K\}\). We consider the DPG\(_e\) method using the lowest order global trial space

\[
U_h = S_h \times Q_h,
\]

where \(Q_h = \{ \hat{\tau} : \hat{\tau}|_{\partial K} \in Q(\partial K) \text{ for all mesh elements } K\} \) and \(S_h = \{w : w|_K \in S(K) \text{ for all mesh elements } K\}\).

Let \(\chi_e\) denote the indicator function of an edge \(e\). If \(a\) denotes a vertex of the square element \(K\), let \(\phi_a\) denote the bilinear function that equals one at \(a\) and equals zero at the other three vertices of \(K\). Let \(\hat{\phi}_a = \phi_a|_{\partial K}\). The collection of eight functions of the form \((0, \hat{\phi}_a)\) and \((\hat{\chi}_e, 0)\), one for each vertex, and one for each edge of \(K\), forms a basis for \(Q(\partial K)\). We distinguish between the horizontal and vertical edges, because the unknowns there approximate different components of the velocity \(\bar{u}\). Accordingly, we will denote by \(\hat{\chi}_h^e\) the indicator function of a horizontal edge and by \(\hat{\chi}_v^e\) the indicator function of a vertical edge.

We now define the local 11 \times 11 DPG matrix for a single element using the basis for \(S(K) \times Q(\partial K)\) obtained by supplementing the basis for \(Q(\partial K)\) described above with the basis for \(S(K)\) consisting of three indicator functions. Enumerating these basis functions as \(e_i, i = 1, \ldots, 11\), the local DPG matrix \(B = B(\omega, \varepsilon)\) is defined by

\[
B_{ij} = b(e_j, T^r e_i),
\]

where \(T^r\) is as defined in (2.11). The basis for the space \(V^r\) is chosen such that each basis function is supported on one element. In our computations, we did not specialize the basis for \(V^r\) any further so, to overcome round-off problems due to ill-conditioned local matrices, we resorted to high precision arithmetic for these local computations.

Given a square element with sides of length \(h\) parallel to the axes, \(B\) can be computed by mapping to the reference element \(\hat{K} = [0, 1]^2\). For any function \(v\) on \(K\) (resp., \(\partial K\)), let the mapped function \(\hat{v}\) on \(\hat{K}\) (resp., \(\partial \hat{K}\)) be defined by \(\hat{v}(\hat{x}) = v(h \hat{x} + \hat{b}_K)\), with \(\hat{b}_K\) such that \(K - \hat{b}_K = h \hat{K}\). The mapped functions \(\hat{e}_i\) are precisely the basis vectors of \(S(\hat{K}) \times Q(\partial \hat{K})\) used when applying (4.1) to compute \(\hat{B}(\omega, \varepsilon)\), the local DPG matrix for \(\hat{K}\). By a change of variables, it is easy to see that

\[
B(\omega, \varepsilon) = h^2 \hat{B}(\omega h, \varepsilon h).
\]

Thus we may compute local DPG matrices by scaling the local DPG matrix for the fixed reference element \(\hat{K}\) obtained using the normalized wavenumber \(\omega h\) and scaling parameter \(\varepsilon h\). It is enough to compute the element matrix \(\hat{B}\) using high precision arithmetic for the ensuing dispersion analysis.

Next, we eliminate the three interior variables of \(S(K)\) and consider the condensed 8 \times 8 local stiffness matrix for the variables in \(Q(\partial K)\). At this stage it will be useful to classify these eight variables (unknowns) into three categories: 1. unknowns at
vertices $a$ (which are the coefficients multiplying the basis function $\hat{\phi}_a$) denoted by “$\bullet$”; 2. unknowns on horizontal edges (coefficients multiplying $\hat{\chi}_e^h$) denoted by “$\uparrow$”; and 3. unknowns on vertical edges (coefficients multiplying the corresponding $\hat{\chi}_e^v$) denoted by “$\rightarrow$”. The normal vectors on all horizontal and vertical edges are fixed to be $(0,1)$ and $(1,0)$, resp., corresponding to the direction of the above-indicated arrows.

Now suppose the mesh is a uniform mesh of congruent square elements. Assembling the above-described condensed $8 \times 8$ element matrices on such a mesh, we obtain a global system where the interior variables are all condensed out. The resulting equations can be represented using the stencils in Figure 4.1. A row of the matrix system corresponding to an unknown of the type $\bullet$ has 21 nonzero entries corresponding to unknowns of all three types, as shown in Figure 4.1(a). Similarly, the unknowns of the type $\uparrow$ and $\rightarrow$ connect to other unknowns in the 13-point stencils depicted in Figures 4.1(b) and 4.1(c), resp. These stencils form the basis of our dispersion analysis in the next section.

5. Dispersion analysis. This section is devoted to a numerical study of the DPG$_\varepsilon$ method, by means of a dispersion analysis. The dispersion analysis is motivated by [13]. Details on dispersion analyses applied to standard methods can be found in [1] and the extensive bibliography presented therein.

5.1. The approach. To briefly adapt the approach of [13] to fit our context, we consider a general method for the homogeneous Helmholtz equation on an infinite uniform lattice $(h\mathbb{Z})^2$. Suppose the method has $S$ different types of nodes on this lattice, some falling in-between the lattice points, each corresponding to a different type of variable, with its own stencil (and hence its own equation). All nodes of the $t$th type $(t = 1, 2, \ldots, S)$ are assumed to be of the form $\vec{j}h$ where $\vec{j}$ lies in an infinite subset of $(\mathbb{Z}/2)^2$. The solution value at a general node $\vec{j}h$ of the $t$th type is denoted by $\psi_{t,\vec{j}}$. Note that methods with multiple solution components are accommodated using the above-mentioned node types.

The $t$th stencil, centered around a node of the $t$th type at position $\vec{j}h$, consists of a finite number of nodes of potentially all types. Suppose we have finite index sets $J_s \subset (\mathbb{Z}/2)^2$, for each $s = 1, 2, \ldots, S$, such that all the nodes of the $t$th stencil centered around $\vec{j}h$ can be listed as $N_{\vec{j},t} = \{(\vec{j} + \vec{l})h : \vec{l} \in J_s \text{ for some } s = 1, 2, \ldots, S\}$ with the understanding that $(\vec{j} + \vec{l})h$ is a node of $s$th type whenever $\vec{l} \in J_s$. This allows interaction between variables of multiple types. Every node $(\vec{j} + \vec{l})h$ in $N_{\vec{j},t}$ has a corresponding stencil coefficient (or weight) denoted by $D_{t,s,\vec{l}}$, which are the...
linear combinations of the entries of $B(\omega, \varepsilon)$ that likewise arise as entries of the global matrix of (2.12). Due to translational invariance, these weights do not change if we center the stencil at another node at position $j' h$. Hence, the numbers $D_{t,s,l}$ do not depend on the center index $j$. We obtain the method’s equation at a general node $j h$ of the $t$th type by applying the $t$th stencil centered around $j h$ to the solution values $\{\psi_{t,j}\}$, namely,

\[
\sum_{s=1}^{S} \sum_{l \in J_s} D_{t,s,l} \psi_{s,j+l} = 0.
\]

Note that we have set all sources to zero to get a zero right-hand side in (5.1).

Plane waves, $\psi(x) \equiv e^{i\vec{k} \cdot \vec{x}}$, are exact solutions of the Helmholtz equation with zero sources (and are often used to represent other solutions). Here the wave vector $\vec{k}$ is of the form $\vec{k} = (\omega, \sin(\theta))$ for some $0 \leq \theta < 2\pi$ representing the direction of propagation. The objective of dispersion analysis is to find similar solutions of the discrete homogeneous system. Accordingly, we set in (5.1) the ansatz

\[
\psi_{t,j} = a_t e^{i\vec{k}_t \cdot \vec{h}},
\]

where $\vec{k}_t = \omega h (\cos(\theta), \sin(\theta))$ and $a_t$ is an arbitrary complex number associated with the $t$th variable type. We want to find such discrete wavenumbers $\omega h$ satisfying (5.1).

To this end, we must solve (5.1) after substituting (5.2) therein, namely,

\[
\sum_{s=1}^{S} a_s \sum_{l \in J_s} D_{t,s,l} e^{i\vec{k}_s \cdot (j+l)h} = 0
\]

for all $t = 1, 2, \ldots S$. Multiplying by $e^{-i\vec{k}_h \cdot \vec{j} h}$, we remove any dependence on $\vec{j}$. Defining the $S \times S$ matrix $F(\omega h)$ by

\[
F_{ts}(\omega h) = \sum_{l \in J_s} D_{t,s,l} e^{i(\omega h (\cos \theta, \sin \theta) \cdot l) h},
\]

we observe that solving (5.3) is equivalent to solving

\[
\det F(\omega h) = 0.
\]

This is the nonlinear equation we solve to obtain the discrete wavenumber $\omega h$ corresponding to any given $\theta$ and $\omega$.

### 5.2. Application to the DPG$_{\varepsilon}$ method.

Next, we apply the above-described framework to the lowest order DPG stencil discussed in section 4. Since there are three different types of stencils (see Figure 4.1), we have $S = 3$. Unknowns of the first type, denoted by $\bullet$, represent the DPG method’s approximation to the value of $\phi$ at the nodes $j h$ for all $j \in Z^2$. The stencil of the first type is the one shown in Figure 4.1(a). The unknowns of the second type represent the method’s approximation to the vertical components of $\vec{u}$ on the midpoints of horizontal edges, i.e., at all points in $(hZ + h/2) \times hZ$. These unknowns are denoted by $\uparrow$ and have the stencil portrayed in Figure 4.1(b). Similarly, the third type of stencil and unknown are as in Figure 4.1(c).
To summarize, (5.2) in the lowest order DPG case, becomes

\[
\begin{align*}
\psi_{1,j} &= \hat{\phi}_h(x_j) = a_1 e^{i\kappa_h \cdot x_j} \quad \forall x_j \\
\psi_{2,j} &= \hat{u}_h(x_j) = a_2 e^{i\kappa_h \cdot x_j} \quad \forall x_j \\
\psi_{3,j} &= \hat{u}_h(x_j) = a_3 e^{i\kappa_h \cdot x_j} \quad \forall x_j
\end{align*}
\]

The condensed 8 × 8 DPG matrices, discussed in section 4, can be used to compute the stencil weights \(D_{s,t} \) in each of the three cases, which in turn lead to the nonlinear equation (5.4) for any given propagation angle \(\theta\).

We numerically solved the nonlinear equation for \(\omega_h\), for various choices of \(\theta\) (propagation angle), \(r\) (enrichment degree), \(\varepsilon\) (scaling factor in the \(V\)-norm), and \(h\) (mesh size). The first important observation from our computations is that the computed wavenumbers \(\omega_h\) are complex numbers. They lie close to \(\omega\) in the complex plane. The small but nonzero imaginary parts of \(\omega_h\) indicate that the DPG \(\varepsilon\) method has dissipation errors, in addition to dispersion errors. The results are described in more detail below.

5.3. Dependence on \(\theta\). To understand how dispersion errors vary with propagation angle \(\theta\), we fix the exact wavenumber \(\omega\) appearing in the Helmholtz equation to 1 (so the wavelength is 2\(\pi\)) and examine the computed \(\text{Re}(\omega_h)\) for each \(\theta\).

One way to visualize the results is through a plot of the corresponding discrete wavevectors \(\text{Re}(\kappa_h)\) versus \(\kappa\) for every propagation direction \(\theta\). Due to symmetry, we only need to examine this plot in the region \(0 \leq \theta \leq \pi/2\). We present these plots for the case \(r = 3\) in Figure 5.1. We fix \(h = 2\pi/4\). (This corresponds to four elements per wavelength if the propagation direction is aligned with a coordinate axis.) In Figure 5.1(a), we plot the curve traced out by the endpoints of the discrete wavevectors \(\kappa_h\). We see that as \(\varepsilon\) decreases, the curve gets closer to the (solid) circle traced out by the exact wavevector \(\kappa\). This indicates better control of dispersive errors with decreasing \(\varepsilon\) (cf. Theorem 3.1).

In Figure 5.1(b), we compare the \(\kappa_h\) obtained using the lowest order DPG method with the discrete wavenumbers of the standard lowest order (bilinear) finite element
method (FEM). Clearly the wave speeds obtained from the DPG method are closer to the exact \( \omega = 1 \) than those obtained by bilinear FEM. However, since the lowest order DPG method has a larger stencil than bilinear FEM, one may argue that a better comparison is with methods having the same stencil size. We therefore compare the DPG method with two other methods which have stencils of exactly the same shape and size: (i) The biquadratic FEM (which after condensation has three stencils of the same size as the lowest order DPG method), and (ii) the conforming first order \( L^2(\Omega) \) least-squares method using the lowest order Raviart–Thomas and Lagrange spaces (which has no interior nodes to condense out). While the wave speeds from the DPG method did not compare favorably with the biquadratic FEM of (i), we found that the DPG method performs better than the least-squares method in (ii).

### 5.4. Dependence on \( \varepsilon \) and \( r \)

We have seen in Figure 5.1 that the discrete wave speed \( \omega_h \) is a function of the propagation angle \( \theta \). We now examine the maximum discrepancy between real and imaginary parts of \( \omega_h \) and \( \omega \) over all angles. Accordingly, define

\[
\rho = \max_{\theta} |\text{Re}(\omega_h(\theta)) - \omega|, \quad \eta = \max_{\theta} |\text{Im}(\omega_h(\theta))|.
\]

The former indicates dispersive errors while the latter indicates dissipative errors. Fixing \( \omega = 1 \) and \( h = 2\pi/8 \) (so that there are about eight elements per wavelength), we examine these quantities as a function of \( r \) and \( \varepsilon \) in Figure 5.2. The first of the plots in Figures 5.2(a) and 5.2(b) show that the errors decrease as \( \varepsilon \) decreases from 1 to about 0.1. In view of Theorem 3.1, we expected this decrease.

However, the behavior of the method for smaller \( \varepsilon \) is curious. In the remaining plots of Figure 5.2 we see that when \( r \) is odd, the errors continue to decrease for smaller \( \varepsilon \), while for even \( r \), the errors start to increase as \( \varepsilon \to 0 \). We do not yet understand this enough to give a theoretical explanation.

### 5.5. Dependence on \( \omega \)

Now we examine how \( \omega_h \) depends on \( \omega \). First, let us note that the matrix \( F \) in (5.4) only depends on \( \omega_h \). (This can be seen, for instance, from (4.2) and noting how the stencil weights depend on the entries of \( B \).) Hence, we will study how \( \omega_h \) depends on the normalized wavenumber \( \omega h \), restricting ourselves to the case of \( \theta = 0 \).

In Figure 5.3(a), we plot (in logarithmic scale) the absolute value of \( \omega_h h - \omega h \) versus \( \omega h \) for the standard bilinear FEM, the lowest order \( L^2 \) least-squares method (marked LS), and the DPG method with \( \varepsilon = 10^{-6}, r = 3 \). We observe that while \( |\omega_h h - \omega h| \) appears to decrease at \( O(\omega h)^2 \) for the least-squares method, it appears to decrease at the higher rate of \( O(\omega h)^3 \) for the FEM and DPG cases considered in the same graph. For easy reference, we have also plotted lines indicating slopes corresponding to \( O(\omega h)^2 \) and \( O(\omega h)^3 \) decrease, marked "quadratic" and "cubic", resp., in the same graph.

Note that a convergence rate of \( |\omega_h h - \omega h| = O(\omega h)^3 \) implies that the difference between discrete and exact wave speeds goes to zero at the rate

\[
|\omega_h - \omega| = \omega O(\omega h)^2.
\]

This shows the presence of the so-called [2] pollution errors: For instance, as \( \omega \) increases, even if we use finer meshes so as to maintain \( \omega h \) fixed, the error in wave speeds will continue to grow at the rate of \( O(\omega) \). Our results show that pollution errors are present in all three methods considered in Figure 5.3(a). The difference
The discrepancies between exact and discrete wavenumbers as a function of $\varepsilon$, when $\omega = 1$ and $h = 2\pi/8$.

Fig. 5.2.

(a) Plot of $|\omega_h h - \omega| h$ for three methods

(b) Case of DPG with $r = 3$ and various $\varepsilon$

Fig. 5.3. Rates of convergence of $|\omega_h h - \omega| h$ to zero for small $\omega h$, in the case of propagation angle $\theta = 0$. 

Comparison of three methods

Convergence of $\omega h$
in convergence rates, e.g., whether $|\omega_h - \omega|$ converges to zero at the rate $\omega O(\omega h)^2$ or at the rate $\omega O(\omega h)$, becomes important, for example, when trying to answer the following question: What $h$ should we use to obtain a fixed error bound for $|\omega_h - \omega|$ for all frequencies $\omega$? While methods with convergence rate $\omega O(\omega h)$ would require $h \approx \omega^{-2}$, methods with convergence rate $\omega O(\omega h)^2$ would only require $h \approx \omega^{-3/2}$.

Next, consider Figure 5.3(b), where we observe interesting differences in convergence rates within the $\text{DPG}_\varepsilon$ family. While the $\text{DPG}_\varepsilon$ method for $\varepsilon = 1$ exhibits the same quadratic rate of convergence as the least-squares method, we observe that a transition to higher rates of convergence progressively occurs as $\varepsilon$ is decreased by each order of magnitude. The $\varepsilon = 10^{-6}$ case shows a rate virtually indistinguishable from the cubic rate in the considered range. The convergence behavior of the $\text{DPG}_\varepsilon$ method thus seems to vary “in-between” those of the least-squares method and the standard FEM as $\varepsilon$ is decreased. The values of $\omega h$ considered in these plots are $2\pi/2^l$ for $l = 1, 2, \ldots, 7$, which cover the numbers of elements per wavelength in usual practice.

Next, we consider a wider range of $\omega h$ following [21], where such a study was done for standard finite elements, separating the real and imaginary parts of $\omega_h$. Our results for the case of $\theta = 0$ are collected in Figure 5.4. To discuss these results, let us first recall the behavior of the standard bilinear FEM (whose discrete wavenumbers are also plotted in the dash-dotted curves in Figure 5.4). From its well-known dispersion relation (see, e.g., [1]), we observe that if $\omega_h$ solves the dispersion relation, then $2\pi - \omega_h$ also solves it. Accordingly, the plot in Figure 5.4(a) is symmetric about the horizontal line at height $\pi$. Furthermore, as already shown in [21], $\omega_h$ is real valued in the range $0 < \omega_h < \sqrt{12}$. The threshold value $\omega h = \sqrt{12}$ was called the “cut-off” frequency. (Note that in the regime $\omega h > \pi$, we have less than two elements per wavelength. Note also that $\sqrt{12} > \pi$.) As can be seen from Figures 5.4(a) and 5.4(b), in the range $\sqrt{12} < \omega h \leq 6$, the bilinear finite elements yield $\omega_h$ with a constant real part of $\pi$ and nonzero imaginary parts of increasing magnitude.

We observed a somewhat similar behavior for the $\text{DPG}_\varepsilon$ method—see the solid curves of Figure 5.4, which were obtained after calculating $F$ explicitly using the
computer algebra package Maple, for the lowest order DPG\textsubscript{ε} method, setting \( r = 3 \) and \( \varepsilon = 0 \). The major difference between the DPG\textsubscript{ε} and FEM results is that \( \omega_h h \) from the DPG\textsubscript{ε} method was not real valued even in the regime where FEM wavenumbers were real. It seems difficult to define any useful analogue of the cut-off frequency in this situation. Nonetheless, we observe from Figures 5.4(a) and 5.4(b) that there is a segment of constant real part of value \( \pi \), before which the imaginary part of \( \omega_h h \) is relatively small. As the number of mesh elements per wavelength increases (i.e., as \( \omega h \) becomes smaller), the imaginary part of \( \omega_h h \) becomes small. We therefore expect the diffusive errors in the DPG\textsubscript{ε} method to be small when \( \omega h \) is small. Finally, we also conclude from Figure 5.4 that both the dispersive and dissipative errors are better behaved for the DPG\textsubscript{ε} method when compared to the \( L^2 \) least-squares method.

6. Conclusions. The DPG\textsubscript{ε} method for the Helmholtz equation was presented and analyzed. The case \( \varepsilon = 1 \) was analyzed previously in [12]. The numerical results in [12] showed that in a comparison of the ratio of \( L^2 \)-norms of the discretization error to the best approximation error, the DPG method had superior properties. The pollution errors reported in [12] for a higher order DPG method were so small that its growth could not be determined conclusively there. In this paper, by performing a dispersion analysis on the DPG method for the lowest possible order, we found that the method has pollution errors that asymptotically grow with \( \omega \) at the same rate as other comparable methods.

In addition, we found both dispersive and dissipative types of errors in the lowest order DPG method. The dissipative errors manifest in computed solutions as artificial damping of wave amplitudes (e.g., as illustrated in Figure 1.1).

Our results show that the DPG solutions have higher accuracy than an \( L^2 \)-based least-squares method with a stencil of identical size. However, the errors in the (lowest order) DPG method did not compare favorably with a standard (higher order) FEM having a stencil of the same size. Whether this disadvantage can be offset by the other advantages of the DPG methods (such as the regularizing effect of \( \varepsilon \), and the fact that it yields Hermitian positive definite linear systems and good gradient approximations) remains to be investigated.

We provided the first theoretical justification for considering the \( \varepsilon \)-modified DPG method. If the test space were exactly computed, then Theorem 3.1 shows that the errors in numerical fluxes and traces will improve as \( \varepsilon \to 0 \). However, if the test space is inexactly computed using the enrichment degree \( r \), then the numerical results from the dispersion analysis showed that errors continually decreased as \( \varepsilon \) was decreased only for odd \( r \). A full theoretical explanation of such discrete effects and the limiting behavior when \( \varepsilon \) is 0 deserve further study.

REFERENCES

DISPERSION ANALYSIS FOR DPG METHODS

A39


