ANALYSIS OF FEAST SPECTRAL APPROXIMATIONS USING THE DPG DISCRETIZATION

JAY GOPALAKRISHNAN, LUKA GRUBIŠIĆ, JEFFREY OVALL, AND BENJAMIN Q. PARKER

ABSTRACT. A filtered subspace iteration for computing a cluster of eigenvalues and its accompanying eigenspace, known as "FEAST", has gained considerable attention in recent years. This work studies issues that arise when FEAST is applied to compute part of the spectrum of an unbounded partial differential operator. Specifically, when the resolvent of the partial differential operator is approximated by the discontinuous Petrov Galerkin (DPG) method, it is shown that there is no spectral pollution. The theory also provides bounds on the discretization errors in the spectral approximations. Numerical experiments for simple operators illustrate the theory and also indicate the value of the algorithm beyond the confines of the theoretical assumptions. The utility of the algorithm is illustrated by applying it to compute guided transverse core modes of a realistic optical fiber.

1. INTRODUCTION

We study certain numerical approximations of the eigenspace associated to a cluster of eigenvalues of a reaction-diffusion operator, namely the unbounded operator $A = -\Delta - \nu$ in $L^2(\Omega)$, whose domain is $H_0^1(\Omega)$. Here $\nu \in L^\infty(\Omega)$ and $\Omega \subset \mathbb{R}^n$ is an open bounded set with Lipschitz boundary. The eigenvalue cluster of interest is assumed to be contained inside a finite contour Γ in the complex plane $\mathbb{C}$. The computational technique is the FEAST algorithm [16], which is now well known as a subspace iteration, applied to an approximation of an operator-valued contour integral over Γ. This technique requires one to approximate the resolvent function $z \mapsto R(z) = (z - A)^{-1}$ at a few points along the contour. The specific focus of this paper is the discretization error in the final spectral approximations when the discontinuous Petrov Galerkin (DPG) method [7] is used to approximate the resolvent.

Contour integral methods [1, 13, 16, 21], such as FEAST, have been gaining popularity in numerical linear algebra. When used as an algorithm for matrix eigenvalues, discretization errors are irrelevant, which explains the dearth of studies on discretization errors within such algorithms. However, in this paper, like in [9, 14], we are interested in the eigenvalues of a partial differential operator on an infinite-dimensional space. In these cases, practical computations can proceed only after discretizing the resolvent of the partial differential operator by some numerical strategy, such as the finite element method. We specifically focus on the DPG method, a least-squares type of finite element method.

One of our motivations for considering the DPG discretization is that it allows us to approximate $R(z)$ by solving a sparse Hermitian positive definite system (even when $z - A$ is indefinite) using efficient iterative solvers. Another practical reason is that it offers a built-in (a posteriori) error estimator in the resolvent approximation (see [4]), thus immediately

suggesting a straightforward algorithmic avenue for eigenspace error control. The exploitation of these advantages, including the design of preconditioners and adaptive algorithms, are postponed to future work. The focus of this paper is limited to obtaining *a priori* error bounds and convergence rates for the computed eigenspace and accompanying Ritz values.

According to [9], bounds on spectral errors can be obtained from bounds on the approximation of the resolvent $z \mapsto R(z)$. This function maps complex numbers to bounded operators. In [9], certain finite-rank computable approximations to $R(z)$, denoted by $R_h(z)$, were considered and certain abstract sufficient conditions were laid out for bounding the resulting spectral errors. (Here h represents some discretization parameter like the mesh size.) This framework is summarized in Section 2. Our approach to the analysis in this paper is to verify the conditions of this abstract framework when $R_h(z)$ is obtained using the DPG discretization.

One of our applications of interest is the fast and accurate computation of the guided modes of optical fibers. In the design and optimization of new optical fibers, such as the emerging microstructured fibers, one often needs to compute such modes many hundreds of times for varying parameters. FEAST appears to offer a well-suited method for this purpose. The Helmholtz operator arising from the fiber eigenproblem is of the above-mentioned type (wherein ν is related to the fiber's refractive index). In Section 5, we will show the efficacy of the FEAST algorithm, combined with the DPG resolvent discretization, by computing the modes of a commercially marketed step-index fiber.

The outline of the paper is as follows. In Section 2 we present the abstract theory from [9] pertaining to FEAST iterations using discretized resolvents of unbounded operators. In Section 3 we derive new estimates for discretizations of a resolvent by the DPG method. In Section 4 we present benchmark results on problems with well-known solutions which serve as a validation of the method. Finally, in Section 5 we apply the method to compute the modes of a ytterbium-doped optical fiber.

2. The abstract framework

In this section, we summarize the abstract framework of [9] for analyzing spectral discretization errors of the FEAST algorithm when applied to general selfadjoint operators. Accordingly, in this section, A is not restricted to the reaction-diffusion operator mentioned in Section 1. Here we let A be a linear, closed, selfadjoint (possibly unbounded) operator $A : \mathrm{dom}(A) \subseteq \mathcal{H} \to \mathcal{H}$ in a complex Hilbert space $\mathcal{H}$, whose real spectrum is denoted by $\Sigma(A)$. We are interested in approximating a subset $\Lambda \subset \Sigma(A)$ that consists of a finite collection of eigenvalues of finite multiplicity, as well as its associated eigenspace E (the span of all the eigenvectors associated with elements of Λ).

The FEAST iteration uses a rational function

$$(1) \qquad r_N(\xi) = w_N + \sum_{k=0}^{N-1} w_k (z_k - \xi)^{-1} \ .$$

Here the choices of $w_k, z_k \in \mathbb{C}$ are typically motivated by quadrature approximations of the Dunford-Taylor integral

$$(2) \qquad S = \frac{1}{2\pi \mathrm{i}} \oint_\Gamma R(z) \, dz,$$

where $R(z) = (z - A)^{-1}$ denotes the resolvent of A at z. Above, Γ is a positively oriented, simple, closed contour Γ that encloses Λ and excludes $\Sigma(A)\backslash\Lambda$, so that S is the exact spectral projector onto E. Define

$$S_N = r_N(A) = w_N + \sum_{k=0}^{N-1} w_k R(z_k).$$

More details on examples of r_N and their properties can be found in [13, 16].

We are particularly interested in a further approximation of S_N given by

$$(3) \qquad S_N^h = w_N + \sum_{k=0}^{N-1} w_k R_h(z_k).$$

Here $R_h(z) : \mathcal{H} \to \mathcal{V}_h$ is a finite-rank approximation of the resolvent $R(z)$, $\mathcal{V}_h$ is a finite-dimensional subspace of a complex Hilbert space $\mathcal{V}$ embedded in $\mathcal{H}$, and h is a parameter inversely related to $\dim(\mathcal{V}_h)$ such as a mesh size parameter. Note that there is no requirement that these resolvent approximations are such that S_N^h is selfadjoint. In fact, as we shall see later (see Remark 3.4), the S_N^h generated by the DPG approximation of the resolvent is not generally selfadjoint.

We consider a version of the FEAST iterations that use the above approximations. Namely, starting with a subspace $E_h^{(0)} \subseteq \mathcal{V}_h$, compute

$$(4) \qquad E_h^{(\ell)} = S_N^h E_h^{(\ell-1)}, \qquad \text{for } \ell = 1, 2, \dots.$$

If A is a selfadjoint operator on a finite-dimensional space $\mathcal{H}$ (such as the one given by a Hermitian matrix), then one may directly use S_N instead of S_N^h in (4). This case is the well-studied FEAST iteration for Hermitian matrices, which can approximate spectral clusters of A that are strictly separated from the remainder of the spectrum. In our abstract framework for discretization error analysis, we place a similar separation assumption on the exact undiscretized spectral parts Λ and $\Sigma(A)\backslash\Lambda$. Consider the following strictly separated sets $I_\gamma^y = \{x \in \mathbb{R} : |x - y| \leqslant \gamma\}$, and $O_{\delta,\gamma}^y = \{x \in \mathbb{R} : |x - y| \geqslant (1+\delta)\gamma\}$, for some $y \in \mathbb{R}$, $\delta > 0$ and $\gamma > 0$. Using these sets and the quantities

$$(5) \qquad W = \sum_{k=0}^{N} |w_k|, \qquad \hat{\kappa} = \frac{\sup\limits_{x \in O_{\delta,\gamma}^y} |r_N(x)|}{\inf\limits_{x \in I_\gamma^y} |r_N(x)|}.$$

we formulate a spectral separation assumption below.

Assumption 1. There are $y \in \mathbb{R}$, $\delta > 0$ and $\gamma > 0$ such that

$$(6) \qquad \Lambda \subset I_\gamma^y, \qquad \Sigma(A)\backslash\Lambda \subset O_{\delta,\gamma}^y,$$

and that r_N is a rational function of the form (1) with the following properties:

$$z_k \notin \overline{\Sigma(A)}, \quad W < \infty, \quad \text{and} \quad \hat{\kappa} < 1.$$

Assumption 2. The Hilbert space $\mathcal{V}$ is such that $E \subseteq \mathcal{V} \subseteq \mathcal{H}$, there is a $C_\mathcal{V} > 0$ such that for all $u \in \mathcal{V}$, $\|u\|_\mathcal{H} \leqslant C_\mathcal{V}\|u\|_\mathcal{V}$, and $\mathcal{V}$ is an invariant subspace of $R(z)$ for all z in the resolvent set of A, i.e., $R(z)\mathcal{V} \subseteq \mathcal{V}$. (We allow $\mathcal{V} = \mathcal{H}$, and further examples where $\mathcal{V} \neq \mathcal{H}$ can be found in [9, §2].)

Assumption 3. The operators $R_h(z_k)$ and $R(z_k)$ are bounded in $\mathcal{V}$ and satisfy

$$
(7) \qquad \lim_{h \to 0} \max_{k=0,\dots,N-1} \| R_h(z_k) - R(z_k) \|_{\mathcal{V}} = 0.
$$

Assumption 4. Assume that $\mathcal{V}_h$ is contained in $\mathrm{dom}(a)$, where $a(\cdot,\cdot)$ denotes the symmetric (possibly unbounded) sesquilinear form associated to the operator A (as described in, say, [19, §10.2] or [9, §5]).

Various examples of situations where one or more of these assumptions hold can be found in [9]. Next, we proceed to describe the main consequences of these assumptions of interest here. Let $\Lambda = \{\lambda_1, \dots, \lambda_m\}$, counting multiplicities, so that $m = \dim(E)$. By the strict separation of Assumption 1, we can find a curve Θ that encloses $\mu_i = r_N(\lambda_i)$ and no other eigenvalues of S_N. By Assumption 3, S_N^h converges to S_N in norm, so for sufficiently small h, the integral

$$
P_h = \frac{1}{2\pi \mathrm{i}} \oint_{\Theta} (z - S_N^h)^{-1} \, dz
$$

is well defined and equals the spectral projector of S_N^h associated with the contour Θ. Let E_h denote the range of P_h. Now, let us turn to the iteration (4). We shall tacitly assume throughout this paper that $E_h^{(0)} \subseteq \mathcal{V}_h$ is chosen so that $\dim E_h^{(0)} = \dim(P_h E_h^{(0)}) = m$. In practice, this is not restrictive: we usually start with a larger than necessary $E_h^{(0)}$ and truncate it to dimension m as the iteration progresses.

In order to describe convergence of spaces, we need to measure the distance between two linear subspaces M and L of $\mathcal{V}$. For this, we use the standard notion of gap [15] defined by

$$
(8) \qquad \mathrm{gap}_{\mathcal{V}}(M, L) = \max \left[\sup_{m \in U_M^{\mathcal{V}}} \mathrm{dist}_{\mathcal{V}}(m, L), \ \sup_{l \in U_L^{\mathcal{V}}} \mathrm{dist}_{\mathcal{V}}(l, M) \right],
$$

where $\mathrm{dist}_{\mathcal{V}}(x, S) = \inf_{s \in S} \| x - s \|_{\mathcal{V}}$ and $U_M^{\mathcal{V}}$ denotes the unit ball $\{ w \in M : \| w \|_{\mathcal{V}} = 1 \}$ of M. The set of approximations to Λ is defined by

$$
\Lambda_h = \{ \lambda_h \in \mathbb{R} : \ \exists 0 \neq u_h \in E_h \text{ satisfying } a(u_h, v_h) = \lambda_h(u_h, v_h) \text{ for all } v_h \in E_h \}.
$$

In other words, Λ_h is the set of Ritz values of the compression of A on E. The sets Λ and Λ_h are compared using the Hausdorff distance. We recall that the Hausdorff distance between two subsets $\Upsilon_1, \Upsilon_2 \subset \mathbb{R}$ is defined by

$$
\mathrm{dist}(\Upsilon_1, \Upsilon_2) = \max \left[\sup_{\mu_1 \in \Upsilon_1} \mathrm{dist}(\mu_1, \Upsilon_2), \ \sup_{\mu_2 \in \Upsilon_2} \mathrm{dist}(\mu_2, \Upsilon_1) \right],
$$

where $\mathrm{dist}(\mu, \Upsilon) = \inf_{\nu \in \Upsilon} |\mu - \nu|$ for any $\Upsilon \subset \mathbb{R}$. Finally, let C_E denote any finite positive constant satisfying $a(e_1, e_2) \leqslant C_E \| e_1 \|_{\mathcal{H}} \| e_2 \|_{\mathcal{H}}$ for all $e_1, e_2 \in E$. We are now ready to state collectively the following results proved in [9].

Theorem 2.1. *Suppose Assumptions 1–3 hold. Then there are constants $C_N, h_0 > 0$ such that, for all $h < h_0$,*

$$
(9) \qquad \lim_{\ell \to \infty} \mathrm{gap}_{\mathcal{V}}(E_h^{(\ell)}, E_h) = 0,
$$

$$
(10) \qquad \lim_{h \to 0} \mathrm{gap}_{\mathcal{V}}(E, E_h) = 0,
$$

$$
(11) \qquad \mathrm{gap}_{\mathcal{V}}(E, E_h) \leqslant C_N W \max_{k=0,\dots,N-1} \left\| \left[R(z_k) - R_h(z_k) \right] \big|_E \right\|_{\mathcal{V}}.
$$

If, in addition, Assumption 4 holds and $\|u\|_{\mathcal{V}} = \||A|^{1/2}u\|_{\mathcal{H}}$, then there are $C_1, h_1 > 0$ such that for all $h < h_1$,

$$(12) \qquad \operatorname{dist}(\Lambda, \Lambda_h) \leqslant (\Lambda_h^{\max})^2 \operatorname{gap}_{\mathcal{V}}(E, E_h)^2 + C_1 C_E \operatorname{gap}_{\mathcal{H}}(E, E_h)^2,$$

where $\Lambda_h^{\max} = \sup_{e_h \in E_h} \||A|^{1/2}e_h\|_{\mathcal{H}}/\|e_h\|_{\mathcal{H}}$ satisfies $(\Lambda_h^{\max})^2 \leqslant [1 - \operatorname{gap}_{\mathcal{V}}(E, E_h)]^{-2} C_E$.

3. Application to a DPG discretization

In this section, we apply the abstract framework of the previous section to obtain convergence rates for eigenvalues and eigenspaces when the DPG discretization is used to approximate the resolvent of a model operator.

3.1. The Dirichlet operator. Throughout this section, we set $\mathcal{H}, \mathcal{V}$, and A by

$$(13) \qquad \mathcal{H} = L^2(\Omega), \quad A = -\Delta, \quad \operatorname{dom}(A) = \{\psi \in H_0^1(\Omega) : \Delta\psi \in L^2(\Omega)\}, \quad \mathcal{V} = H_0^1(\Omega),$$

where $\Omega \subset \mathbb{R}^n$ ($n \geqslant 2$) is a bounded polyhedral domain with Lipschitz boundary. We shall use standard notations for norms ($\|\cdot\|_X$) and seminorms ($|\cdot|_X$) on Sobolev spaces (X). It is easy to see [9] that Assumption 2 holds with these settings. Note that the operator A in (13) is the operator associated to the form

$$a(u, v) = \int_{\Omega} \operatorname{grad} u \cdot \operatorname{grad} \overline{v} \, dx, \quad u, v \in \operatorname{dom}(a) = \mathcal{V} = H_0^1(\Omega)$$

and that the norm $\|u\|_{\mathcal{V}}$, due to the Poincaré inequality, is equivalent to $\||A|^{1/2}u\|_{\mathcal{H}} = \|A^{1/2}u\|_{\mathcal{H}} = \|\operatorname{grad} u\|_{L^2(\Omega)} = |u|_{H^1(\Omega)}$.

The solution of the operator equation $(z - A)u = v$ yields the application of the resolvent $u = R(z)v$. The weak form of this equation may be stated as the problem of finding $u \in H_0^1(\Omega)$ satisfying

$$(14) \qquad b(u, w) = (v, w)_{\mathcal{H}} \qquad \text{for all } w \in H_0^1(\Omega),$$

where

$$b(w_1, w_2) = z(w_1, w_2)_{\mathcal{H}} - a(w_1, w_2)$$

for any $w_1, w_2 \in H_0^1(\Omega)$. As a first step in the analysis, we obtain an inf-sup estimate and a continuity estimate for b. In the ensuing lemmas z is tacitly assumed to be in the resolvent set of A.

Lemma 3.1. *For all $v \in H_0^1(\Omega)$,*

$$\sup_{y \in H_0^1(\Omega)} \frac{|b(v, y)|}{|y|_{H^1(\Omega)}} \geqslant \beta(z)^{-1}|v|_{H^1(\Omega)},$$

where $\beta(z) = \sup\{|\lambda|/|\lambda - z| : \lambda \in \Sigma(A)\}$.

Proof. Let $v \in H_0^1(\Omega)$ be non-zero, and let $w = \overline{z}R(\overline{z})v$. Then

$$b(s, w) = z(s, v)_{\mathcal{H}}, \qquad \text{for all } s \in H_0^1(\Omega).$$

Choosing $s = v$, it follows immediately that

$$(15) \qquad b(v, v - w) = b(v, v) - z\|v\|_{L^2(\Omega)}^2 = -|v|_{H^1(\Omega)}^2.$$

Moreover, $v - w = (I - \overline{z}R(\overline{z}))v = -A\,R(\overline{z})v$. Recall that the identity $\|AR(z)\|_{\mathcal{H}} = \beta(z)$ holds [15, p. 273, Equation (3.17)] for any z in the resolvent set of A. Since $|s|_{H^1(\Omega)} = \|A^{1/2}s\|_{\mathcal{H}}$

for all $s \in H_0^1(\Omega) = \mathrm{dom}(a) = \mathrm{dom}(A^{1/2})$, and since $A^{1/2}$ commutes with $AR(z)$, we conclude that

$$(16) \qquad |v - w|_{H^1(\Omega)} = |AR(\bar{z})v|_{H^1(\Omega)} = \|AR(\bar{z})A^{1/2}v\|_{\mathcal{H}} \leqslant \beta(\bar{z})\|A^{1/2}v\|_{\mathcal{H}} = \beta(z)|v|_{H^1(\Omega)} \, ,$$

where $\beta(\bar{z}) = \beta(z)$ because the spectrum is real. It follows from (15) and (16) that

$$\sup_{y \in H_0^1(\Omega)} \frac{|b(v, y)|}{|y|_{H^1(\Omega)}} \geqslant \frac{|b(v, v - w)|}{|v - w|_{H^1(\Omega)}} \geqslant \frac{|v|_{H^1(\Omega)}^2}{\beta(z)|v|_{H^1(\Omega)}} \, ,$$

as claimed. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square$

3.2. **The DPG resolvent discretization.** We now assume that Ω is partitioned by a conforming simplicial finite element mesh Ω_h. As is usual in finite element theory, while the mesh need not be regular, the shape regularity of the mesh is reflected in the estimates.

To describe the DPG discretization of $z - A$, we begin by introducing the nonstandard variational formulation on which it is based. We will be brief as the method is described in detail in previous works [7, 8]. Define

$$H^1(\Omega_h) = \prod_{K \in \Omega_h} H^1(K), \qquad Q = H(\mathrm{div}, \Omega)/\prod_{K \in \Omega_h} H_0(\mathrm{div}, K),$$

normed respectively by

$$\|v\|_{H^1(\Omega_h)} = \left(\sum_{K \in \Omega_h} \|v\|_{H^1(K)}^2 \right)^{1/2}, \qquad \|q\|_Q = \inf \left\{ \|q - q_0\|_{H(\mathrm{div}, \Omega)} : \, q_0 \in \prod_{K \in \Omega_h} H_0(\mathrm{div}, K) \right\}.$$

On every mesh element K in Ω_h, the trace $q \cdot n|_{\partial K}$ is in $H^{-1/2}(\partial K)$ for any q in $H(\mathrm{div}, K)$. Above, $H_0(\mathrm{div}, K) = \{q \in H(\mathrm{div}, K) : q \cdot n|_{\partial K} = 0\}$. We denote by $\langle q \cdot n, v \rangle_{\partial K}$ the action of this functional on the trace $v|_{\partial K}$ for any v in $H^1(K)$. Next, for any $u \in H_0^1(\Omega)$, $q \in Q$ and $v \in H^1(\Omega_h)$, set

$$b_h((u, q), v) = \sum_{K \in \Omega_h} \left[\langle q \cdot n, \bar{v} \rangle_{\partial K} + \int_K (zu\bar{v} - \mathrm{grad}\, u \cdot \mathrm{grad}\, \bar{v}) \, dx \right].$$

This sesquilinear form gives rise to a well-posed Petrov-Galerkin formulation, as will be clear from the discussion below.

For the DPG discretization, we use the following finite element subspaces. Let L_h denote the Lagrange finite element subspace of $H_0^1(\Omega)$ consisting of continuous functions, which when restricted to any K in Ω_h, are in $P_p(K)$ for some $p \geqslant 1$. Here and throughout, $P_\ell(K)$ denotes the set of polynomials of total degree at most ℓ restricted to K. Note that when applying the earlier abstract framework to the DPG discretization, in addition to (13), we also set

$$(17) \qquad\qquad\qquad\qquad\qquad \mathcal{V}_h = L_h.$$

Let $RT_h \subset H(\mathrm{div}, \Omega)$ denote the well-known Raviart-Thomas finite element subspace consisting of functions whose restriction to any $K \in \Omega_h$ is a polynomial in $P_{p-1}(K)^n + xP_{p-1}(K)$, where x is the coordinate vector. Then we set $Q_h = \{q_h \in Q : q_h|_K \in P_{p-1}(K)^n + xP_{p-1}(K) + H_0(\mathrm{div}, K)\}$. Finally, let $Y_h \subset H^1(\Omega_h)$ consist of functions which, when restricted to any $K \in \Omega_h$, lie in $P_{p+n+1}(K)$.

We now define the approximation of the resolvent action $u = R(z)f$ by the DPG method, denoted by $u_h = R_h(z)f$, for any $f \in L^2(\Omega)$. The function u_h is in L_h. Together with $\varepsilon_h \in Y_h$ and $q_h \in Q_h$, it satisfies

$$(18\text{a}) \qquad (\varepsilon_h, \eta_h)_{H^1(\Omega_h)} + b_h((u_h, q_h), \eta_h) = \int_\Omega f\, \bar{\eta}_h\, dx, \qquad \text{for all } \eta_h \in Y_h,$$

$$(18\text{b}) \qquad b_h((w_h, r_h), \varepsilon_h) = 0, \qquad \text{for all } w_h \in L_h,\ r_h \in Q_h.$$

where

$$(\varepsilon_h, \eta_h)_{H^1(\Omega_h)} = \sum_{K \in \Omega_h} \int_K (\varepsilon_h \bar{\eta}_h + \operatorname{grad} \varepsilon_h \cdot \operatorname{grad} \bar{\eta}_h)\, dx.$$

The distance between u and u_h is bounded in the next result. There and in similar results in the remainder of this section, we tacitly understand z to vary in some bounded subset D of the resolvent set of A in the complex plane (containing the contour Γ) and write $t_1 \lesssim t_2$ whenever there is a positive constant C satisfying $t_1 \leqslant Ct_2$ and C is independent of

$$h = \max_{K \in \Omega_h} \operatorname{diam}(K)$$

but dependent on the diameter of D and the shape regularity of the mesh Ω_h. The deterioration of the estimates as z gets close to the spectrum of A is identified using $\beta(z)$ of Lemma 3.1.

Lemma 3.2. *For all $f \in L^2(\Omega)$,*

$$\|R(z)f - R_h(z)f\|_{\mathcal{V}} \lesssim \beta(z) \left[\inf_{w_h \in L_h} \|u - w_h\|_{H^1(\Omega)} + \inf_{q_h \in RT_h} \|q - q_h\|_{H(\operatorname{div},\Omega)} \right],$$

where $u = R(z)f$ and $q = \operatorname{grad} u$.

Proof. The proof proceeds by verifying the sufficient conditions for convergence of DPG methods known in the existing literature. The result of [11, Theorem 2.1] immediately gives the stated result, provided we verify its three conditions, reproduced below in a form convenient for us. The first two conditions there, taken together, is equivalent to the bijectivity of the operator generated by $b_h(\cdot, \cdot)$. Hence we shall state them in the following alternate form (dual to the form stated in [11]). The first is the uniqueness condition

$$(19\text{a}) \qquad \{\eta \in H^1(\Omega_h) : \ b_h((w, r), \eta) = 0, \ \text{ for all } (w, r) \in H^1_0(\Omega) \times Q\} = \{0\}.$$

The second condition is that there are $C_1, C_2 > 0$ such that

$$(19\text{b}) \qquad C_1 \big[|w|^2_{H^1(\Omega)} + \|r\|^2_Q\big]^{1/2} \leqslant \sup_{\eta \in H^1(\Omega_h)} \frac{|b_h((w, r), \eta)|}{\|\eta\|_{H^1(\Omega_h)}} \leqslant C_2 \big[|w|^2_{H^1(\Omega)} + \|r\|^2_Q\big]^{1/2}$$

for all $w \in H^1_0(\Omega)$ and $r \in Q$. Finally, the third condition is the existence of a bounded linear operator $\Pi_h : H^1(\Omega_h) \to Y_h$ such that

$$(19\text{c}) \qquad b_h((w_h, r_h), \eta - \Pi_h \eta) = 0.$$

Once these conditions are verified, [11, Theorem 2.1] implies

$$(20) \qquad |u - u_h|_{H^1(\Omega)} \leqslant \frac{C_2 \|\Pi\|}{C_1} \left[\inf_{w_h \in L_h} |u - w_h|_{H^1(\Omega)} + \inf_{q_h \in RT_h} \|q - q_h\|_{H(\operatorname{div},\Omega)} \right]$$

with $u = R(z)f$ and $u_h = R_h(z)f$.

It is possible to verify conditions (19a) and (19b) on $b_h(\cdot, \cdot)$ using the properties of $b(\cdot, \cdot)$. First note that [5, Theorem 2.3] shows that

$$\sup_{v \in H^1(\Omega_h)} \frac{|\sum_{K \in \Omega_h} \langle r \cdot n, v \rangle_{\partial K}|}{\|v\|_{H^1(\Omega_h)}} = \|r\|_Q.$$

This, together with [5, Theorem 3.3] implies that the inf-sup condition for b that we proved in Lemma 3.1 implies an inf-sup condition for b_h, namely the lower inequality of (19b) holds with

$$\frac{1}{C_1^2} = \beta(z)^2 + [\beta(z)(1 + |z|) + 1]^2.$$

By combining this with the continuity estimate of b_h with $C_2 = 1 + |z|$, we obtain that C_2/C_1 is $O(\beta(z))$. Finally, Condition (19c) follows from the Fortin operator constructed in [11, Lemma 3.2] whose norm is a constant bounded independently of z. Hence the lemma follows from (20). □

Remark 3.3. Note that the degree of functions in Y_h was chosen to be $p + n + 1$ in order to satisfy the moment condition

$$\int_K (\eta - \Pi_h \eta) w_p \, dx = 0$$

for all $w_p \in P_p(K)$ and $\eta \in H^1(K)$ on all mesh simplices K (see [11]). This moment condition was used while verifying (19c). Other recent ideas, such as those in [2, 6], may be used to reduce Y_h without reducing convergence rates, and thus improve Lemma 3.2 for specific meshes and degrees.

Remark 3.4. The DPG approximation of $u = R(z)f$, given by $u_h = R_h(z)f$, satisfies (18). We may rewrite (18) using $x_h = (u_h, q_h)$,

$$M_h \varepsilon_h + B_h x_h = f_h,$$
$$B_h^* \varepsilon_h = 0.$$

We omit the obvious definitions of operators $B_h : L_h \times Q_h \to Y_h$, $M_h : Y_h \to Y_h$, and that of f_h (an appropriate projection of f). Eliminating ε_h, we find that $u_h = R_h(z)f$ is a component of $x_h = (B_h^* M_h^{-1} B_h)^{-1} B_h^* M_h^{-1} f_h$. Thus, the operator $R_h(z)$ produced by the DPG discretization need not be selfadjoint even when z is on the real line. For the same reason, the filtered operator S_N^h produced by the DPG discretization is *not generally selfadjoint* even when $\{z_k : k = 0, \ldots, N-1\}$ has symmetry about the real line. Note that selfadjointness of S_N^h is not needed in Theorem 2.1 to conclude the convergence of the eigenvalue cluster at double the convergence rate of eigenspace.

3.3. **FEAST iterations with the DPG discretization.** To approximate $E \subseteq \mathcal{V}$, we apply the filtered subspace iteration (4). In this subsection, we complete the analysis of approximation of E by E_h and the accompanying eigenvalue approximation errors. The analysis is an application of the abstract results in Theorem 2.1. To verify the conditions of this theorem, we need some elliptic regularity. This is formalized in the next regularity assumption.

Assumption 5. Suppose there are positive constants C_{reg} and s such that the solution $u^f \in \mathcal{V}$ of the Dirichlet problem $-\Delta u^f = f$ admits the regularity estimate

(21) $$\|u^f\|_{H^{1+s}(\Omega)} \leqslant C_{\text{reg}} \|f\|_{\mathcal{H}} \quad \text{for any } f \in \mathcal{V}.$$

Also suppose that

$$(22) \qquad \|u^f\|_{H^{1+s_E}(\Omega)} \leqslant C_{\text{reg}}\|f\|_{\mathcal{H}} \qquad \text{for any } f \in E.$$

(Since $E \subseteq \mathcal{V}$, (21) implies (22) with s in place of s_E, but in many cases (22) holds with s_E larger than s. This is the reason for additionally assuming (22).)

Its well known that if Ω is convex, Assumption 5 holds with $s = 1$ in (21). If $\Omega \subset \mathbb{R}^2$ is non-convex, with its largest interior angle at a corner being π/α for some $1/2 < \alpha < 1$, Assumption 5 holds with any positive $s < \alpha$. These results can be found in [12], for example.

Lemma 3.5. *Suppose Assumption 5 holds. Then,*

$$(23) \qquad \|R(z)f - R_h(z)f\|_{\mathcal{V}} \lesssim \beta(z)^2 h^{\min(p,s,1)}\|f\|_{\mathcal{V}}, \qquad \text{for all } f \in \mathcal{V},$$

$$(24) \qquad \|R(z)f - R_h(z)f\|_{\mathcal{V}} \lesssim \beta(z)^2 h^{\min(p,s_E)}\|f\|_{\mathcal{V}}, \qquad \text{for all } f \in E.$$

Proof. By Lemma 3.2, the distance between $u = R(z)f$ and $u_h = R_h(z)f$ can be bounded using standard finite element approximation estimates for the Lagrange and Raviart-Thomas spaces, to get

$$(25) \qquad \|u - u_h\|_{H^1(\Omega)} \lesssim \beta(z)\left[h^r|u|_{H^{1+r}(\Omega)} + h^r|q|_{H^r(\Omega)} + h^r|\text{div } q|_{H^r(\Omega)}\right], \qquad \text{for } r \leqslant p,$$

where $q = \text{grad } u$. Note that since u satisfies $b(u,v) = (f,v)_{\mathcal{H}}$ for all $v \in H_0^1(\Omega)$, by Lemma 3.1,

$$(26) \qquad \beta(z)^{-1}|u|_{H^1(\Omega)} \leqslant \sup_{y \in H_0^1(\Omega)} \frac{|b(u,y)|}{|y|_{H_0^1(\Omega)}} = \sup_{y \in H_0^1(\Omega)} \frac{|(f,y)_{\mathcal{H}}|}{|y|_{H_0^1(\Omega)}} = \|f\|_{H^{-1}(\Omega)}.$$

which implies, by the Poincaré inequality,

$$(27) \qquad \|u\|_{\mathcal{H}} \lesssim |u|_{\mathcal{V}} \lesssim \beta(z)\|f\|_{H^{-1}(\Omega)} \lesssim \beta(z)\|f\|_{\mathcal{H}}.$$

Applying elliptic regularity to $\Delta u = f - zu$, for all $r \leqslant s$ and $r \leqslant 1$,

$$|u|_{H^{1+r}(\Omega)} \leqslant C_{\text{reg}}(\|f\|_{\mathcal{H}} + |z|\|u\|_{\mathcal{H}}) \qquad \text{by (21)},$$

$$(28) \qquad\qquad\qquad\quad \lesssim \beta(z)\|f\|_{\mathcal{H}} \qquad\qquad\qquad \text{by (27)},$$

$$(29) \qquad |q|_{H^r(\Omega)} = |\text{grad } u|_{H^r(\Omega)} \lesssim \beta(z)\|f\|_{\mathcal{H}}, \qquad \text{by (28)},$$

$$|\text{div } q|_{H^r(\Omega)} = |f - zu|_{H^r(\Omega)}$$

$$\qquad\qquad\qquad \lesssim |f|_{H^r(\Omega)} + |z|\beta(z)\|f\|_{\mathcal{H}} \qquad \text{by (28)},$$

$$(30) \qquad\qquad\qquad \lesssim \beta(z)\|f\|_{\mathcal{V}} \qquad\qquad\qquad \text{since } r \leqslant 1.$$

Thus for all $0 \leqslant r \leqslant \min(p,s,1)$, using the estimates (28), (29) and (30) in (25), we have proven (23).

The proof of (24) starts off as above using an $f \in E$. But now, due to the potentially higher regularity, we are able to obtain (28) and (29) for $r \leqslant s_E$. Moreover, as in the proof of (30) above, we find that $|\text{div } q|_{H^r(\Omega)} \lesssim \beta(z)\|f\|_{H^r(\Omega)}$. The argument to bound $\|f\|_{H^r(\Omega)}$ by $\|f\|_{\mathcal{V}}$ now requires a slight modification: since $-\Delta f \in E$, the regularity estimate (22) implies $\|f\|_{H^{1+r}(\Omega)} \lesssim \|f\|_{\mathcal{H}}$. Thus

$$|\text{div } q|_{H^r(\Omega)} \lesssim \beta(z)\|f\|_{\mathcal{V}} \qquad \text{for } r \leqslant s_E,$$

i.e., whenever $f \in E$, the estimates (28), (29) and (30) hold for all $0 \leqslant r \leqslant s_E$. Using them in (25), the proof of (24) is complete. $\qquad \square$

Theorem 3.6. *Suppose Assumption 1 (on spectral separation) and Assumption 5 (on elliptic regularity) hold. Then, there are positive constants C_0 and h_0 such that for all $h < h_0$, the FEAST iterates $E_h^{(\ell)}$ obtained using the DPG approximation of the resolvent converge to E_h and*

$$\text{(31)} \qquad \text{gap}_{\mathcal{V}}(E, E_h) \leqslant C_0\, h^{\min(p, s_E)},$$

$$\text{(32)} \qquad \text{dist}(\Lambda, \Lambda_h) \leqslant C_0\, h^{2\min(p, s_E)}.$$

Here C_0 is independent of h (but may depend on $\beta(z_k)^2$, W, C_N, p, Λ, C_{reg}, and the shape regularity of the mesh).

Proof. We apply Theorem 2.1. As we have already noted, Assumption 2 holds for the model Dirichlet problem with the settings in (13). Estimate (23) of Lemma 3.5 verifies Assumption 3. Thus, since Assumptions 1–3 hold, we may now apply (9) of Theorem 2.1 to conclude that $\text{gap}_{\mathcal{V}}(E_h^{(\ell)}, E_h) \to 0$. Moreover, the inequality (11) of Theorem 2.1, when combined with the rate estimate (24) of Lemma 3.5 at each z_k, proves (31).

Finally, to prove (32), noting that the $\mathcal{V}_h$ set in (17) satisfies Assumption 4, we appeal to (12) of Theorem 2.1 to

$$\text{(33)} \qquad \text{dist}(\Lambda, \Lambda_h) \lesssim \text{gap}_{\mathcal{V}}(E, E_h)^2 + \text{gap}_{\mathcal{H}}(E, E_h)^2.$$

To control the last term, first note that $\|e\|_{\mathcal{V}}^2 = a(e, e) \leqslant C_E \|e\|_{\mathcal{H}}^2$ for all $e \in E$. Moreover, by Assumption 2, $\text{dist}_{\mathcal{H}}(e, E_h) \leqslant C_{\mathcal{V}} \text{dist}_{\mathcal{V}}(e, E_h)$. Hence

$$\text{(34)} \qquad \delta_h^{\mathcal{H}} := \sup_{0 \neq e \in E} \frac{\text{dist}_{\mathcal{H}}(e, E_h)}{\|e\|_{\mathcal{H}}} \lesssim \sup_{0 \neq e \in E} \frac{\text{dist}_{\mathcal{V}}(e, E_h)}{\|e\|_{\mathcal{V}}} \leqslant \text{gap}_{\mathcal{V}}(E, E_h).$$

Note that

$$\text{gap}_{\mathcal{H}}(E, E_h) = \max\left[\delta_h^{\mathcal{H}}, \sup_{m \in U_{E_h}^{\mathcal{H}}} \text{dist}_{\mathcal{H}}(m, E)\right].$$

Now, by the already proved estimate of (31), we know that $\text{gap}_{\mathcal{V}}(E, E_h) \to 0$. Hence, when h is sufficiently small, $\text{gap}_{\mathcal{V}}(E, E_h) < 1$, so $\dim(E_h) = \dim(E) = m$. Taking h even smaller if necessary, $\delta_h^{\mathcal{H}} < 1$ by (34), so by [15, Theorem I.6.34], there is a closed subspace $\tilde{E}_h \subseteq E_h$ such that $\text{gap}_{\mathcal{H}}(E, \tilde{E}_h) = \delta_h^{\mathcal{H}} < 1$. But this means that $\dim(\tilde{E}_h) = \dim(E) = \dim(E_h)$, so $\tilde{E}_h = E_h$. Summarizing, for sufficiently small h, we have

$$\text{gap}_{\mathcal{H}}(E, E_h) = \delta_h^{\mathcal{H}} \lesssim \text{gap}_{\mathcal{V}}(E, E_h).$$

Returning to (33), we conclude that

$$\text{dist}(\Lambda, \Lambda_h) \lesssim \text{gap}_{\mathcal{V}}(E, E_h)^2,$$

and the proof is finished using (31). $\qquad \square$

3.4. **A generalization to additive perturbations.** In this short subsection, we will generalize the above theory to the case of the Dirichlet operator when perturbed additively by a real-valued $L^\infty(\Omega)$ reaction term. Let $\nu : \Omega \to \mathbb{R}$ be a function in $L^\infty(\Omega)$ and let

$$\text{(35)} \qquad a(u, v) = \int_\Omega \left[\, \text{grad}\, u \cdot \text{grad}\, \bar{v} - \nu u \bar{v}\,\right] dx$$

for any $u, v \in \text{dom}(a) = \mathcal{V} = H_0^1(\Omega)$. The operator under consideration in this subsection is the unbounded selfadjoint operator A on $\mathcal{H} = L^2(\Omega)$ generated by the form a, per a standard representation theorem [19, Theorem 10.7].

The starting point for our theory in the previous subsections was an inf-sup condition (see Lemma 3.1) for the resolvent form $b(u, v) = z(u, v)_{\mathcal{H}} - a(u, v)$. We claim that Lemma 3.1 can be extended to the new $a(\cdot, \cdot)$. To prove the claim, given any $v \in H_0^1(\Omega)$, we construct a $w \in H_0^1(\Omega)$ slightly differently from the proof of Lemma 3.1, namely

$$w = R(\bar{z})\,(\bar{z}v + \nu v),$$

which solves $b(s, w) = z(s, v)_{\mathcal{H}} + (\nu s, v)_{\mathcal{H}}$ for all $s \in H_0^1(\Omega)$. Then we continue to obtain the identity

$$(36) \qquad b(v, v - w) = -|v|_{H^1(\Omega)}^2.$$

Next, for any $\mu > \|\nu\|_{L^\infty(\Omega)}$, the form domain $\operatorname{dom}(a) = H_0^1(\Omega)$ equals $\operatorname{dom}(A + \mu)^{1/2}$ by [19, Proposition 10.5]. The same result also gives

$$a(u, v) = ((A + \mu)^{1/2}u, (A + \mu)^{1/2}v)_{\mathcal{H}} - \mu(u, v)_{\mathcal{H}} \qquad \text{for all } u, v \in H_0^1(\Omega).$$

Hence

$$(37) \qquad |w|_{H^1(\Omega)}^2 = a(w, w) + (\nu w, w)_{\mathcal{H}} \leqslant a(w, w) + \mu\|w\|_{\mathcal{H}}^2 = \|(A + \mu)^{1/2}w\|_{\mathcal{H}}^2.$$

To proceed, recall that for any z in the resolvent set, functional calculus [3, Theorem 6.4.1] shows that the spectrum of the normal operator $(A+\mu)^{1/2}R(z)$, consists of $\{(\lambda+\mu)^{1/2}/(z-\lambda) : \lambda \in \Sigma(A)\}$. Thus $(A+\mu)^{1/2}R(z)$ is a bounded operator of norm $c_z = \sup\{|\lambda+\mu|^{1/2}/|z-\lambda| : \lambda \in \Sigma(A)\} < \infty$. Hence (37) implies $|w|_{H^1(\Omega)} \leqslant \|(A+\mu)^{1/2}R(\bar{z})\,(\bar{z}v+\nu v)\|_{\mathcal{H}} \leqslant c_z\|\bar{z}v+\nu v\|_{\mathcal{H}}$. Using the Poincaré inequality $c_P\|v\|_{\mathcal{H}} \leqslant |v|_{H^1(\Omega)}$, this implies $|w|_{H^1(\Omega)} \leqslant (|z|+\mu)(c_z/c_P)|v|_{H^1(\Omega)}$, so

$$(38) \qquad |v - w|_{H^1(\Omega)} \leqslant d(z)|v|_{H^1(\Omega)},$$

where $d(z) = 1 + (|z| + \mu)c_z/c_P$. Combining (36) and (38), we have

$$\sup_{y \in H_0^1(\Omega)} \frac{|b(v, y)|}{|y|_{H^1(\Omega)}} \geqslant \frac{|b(v, v - w)|}{|v - w|_{H^1(\Omega)}} \geqslant \frac{|v|_{H^1(\Omega)}^2}{d(z)|v|_{H^1(\Omega)}},$$

so the inf-sup condition follows, extending Lemma 3.1 as claimed.

Lemma 3.7 (Generalization of Lemma 3.1). *Suppose a as in* (35), $b(u, v) = z(u, v)_{\mathcal{H}} - a(u, v)$, *$z$ is in the resolvent set of A, and $d(z)$ is as defined above. Then for all $v \in H_0^1(\Omega)$,*

$$\sup_{y \in H_0^1(\Omega)} \frac{|b(v, y)|}{|y|_{H^1(\Omega)}} \geqslant d(z)^{-1}\,|v|_{H^1(\Omega)}.$$

Using this lemma in place of Lemma 3.1, the remainder of the analysis proceeds with minimal changes, provided we also assume that ν is piecewise constant. More precisely, assume that ν is constant on each element of the mesh Ω_h. Then the same Fortin operator used in the proof of Lemma 3.2 applies. Hence the final result of Theorem 3.6 holds with a possibly different constant C_0 (still independent of h) whenever Assumption 5 holds.

4. Numerical convergence studies

In this section, we report on our numerical convergence studies using the FEAST algorithm with the DPG discretization for the model Dirichlet eigenproblem. This spectral approximation technique is exactly the one described in Section 3.2. An implementation of this technique was built using [10], which contains a hierarchy of Python classes representing approximations of spectral projectors. The DPG discretization is implemented using a python interface into an existing well-known C++ finite element library called NGSolve [20]. We

omit the implementation details of the FEAST algorithm as they can be found either in our public code [10] or previous works like [18, Algorithm 1.1] and [13]. We note that our implementation performs an implicit orthogonalization through a small Rayleigh-Ritz eigenproblem at each iteration. For all experiments reported below, we set r_N to the rational function corresponding to the Butterworth filter obtained by setting $w_N = 0$ and

$$(39) \qquad z_k = \gamma e^{i(\theta_k + \phi)} + y, \qquad w_k = \gamma e^{i(\theta_k + \phi)}/N, \qquad k = 0, \dots, N-1,$$

where $\theta_k = 2\pi k/N$ and $\phi = \pm\pi/N$. This corresponds to an approximation of the contour integral in (2), with a circular contour Γ of radius γ centered at y, using the trapezoidal rule with N equally spaced quadrature points. In all experiments reported below, we set $N = 8$.

4.1. **Discretization errors on the unit square.** Let $\Omega = (0,1) \times (0,1)$ and consider the Dirichlet eigenvalues enclosed within the circular contour Γ of radius $\gamma = 45$ and center $y = 20$. The exact set of eigenvalues for this example is known to be $\Lambda = \{2\pi^2, 5\pi^2\}$. The first eigenvalue $2\pi^2 = \lambda_1$ is of multiplicity 1, while the second $5\pi^2 = \lambda_2 = \lambda_3$ is of multiplicity 2. The corresponding eigenfunctions are well-known analytic functions.

To perform the numerical studies, we begin by solving our problem on a coarse mesh of mesh size $h = 2^{-2}$ and refine until we reach a mesh size of $h = 2^{-7}$. Each mesh refinement halves the mesh size by either bisecting or quadrisecting the triangular elements of a mesh. For each mesh size value of h, we perform this experiment for polynomial degrees $p = 1, 2$, and 3. After each experiment we collect the approximate eigenvalues ordered so that $\lambda_{1,h} \leqslant \lambda_{2,h} \leqslant \lambda_{3,h}$ and their corresponding eigenfunctions $e_{i,h}$.

One way to measure the convergence of eigenfunctions is through

$$\delta_i^{(1)} = \min_{0 \neq e \in E} |e_{i,h} - e|_{H^1(\Omega)} = \text{dist}_{H_0^1(\Omega)}(e_{i,h}, E),$$

$$\delta_i^{(2)} = \min_{0 \neq e_h \in E_h} |e_i - e_h|_{H^1(\Omega)} = \text{dist}_{H_0^1(\Omega)}(e_i, E_h).$$
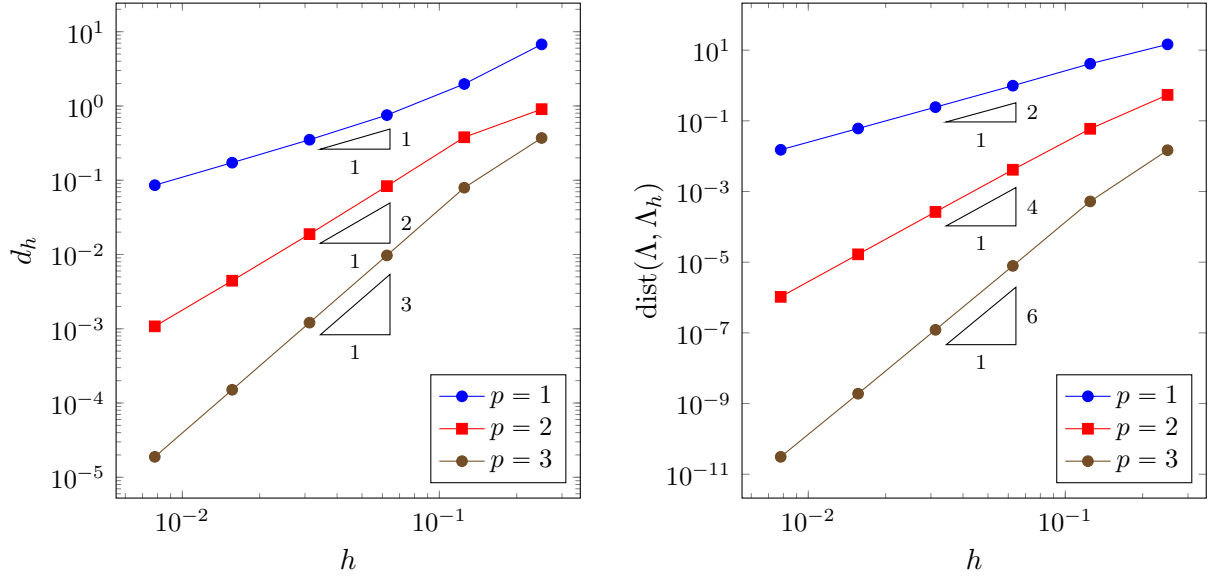
(A) Convergence rates for eigenfunctions   (B) Convergence rates for eigenvalues

FIGURE 1. Results for the unit square

| | λ_1 | | λ_2 | | λ_3 | |
h	ERR	NOC	ERR	NOC	ERR	NOC
2^{-2}	6.29e-02	—	3.29e-02	—	5.95e-02	—
2^{-3}	2.41e-02	1.39	2.65e-03	3.63	4.05e-03	3.88
2^{-4}	9.48e-03	1.34	2.55e-04	3.38	2.59e-04	3.97
2^{-5}	3.75e-03	1.34	2.99e-05	3.09	1.63e-05	3.99
2^{-6}	1.49e-03	1.34	4.03e-06	2.89	1.02e-06	4.00

TABLE 1. Eigenvalue errors (ERR) and numerical order of convergence (NOC) for the smallest three eigenvalues on the L-shaped domain.

Note that both $\delta_i^{(1)}$ and $\delta_i^{(2)}$ are bounded by $\mathrm{gap}_{H_0^1(\Omega)}(E_h, E)$. Since computing $\delta_i^{(1)}$ and $\delta_i^{(2)}$ require exact integration of quantities involving the exact eigenspace, we instead compute

$$\delta_{i,h}^{(1)} = \mathrm{dist}_{H_0^1(\Omega)}(e_{i,h}, I_h E) \quad \text{and} \quad \delta_{i,h}^{(2)} = \mathrm{dist}_{H_0^1(\Omega)}(I_h e_i, E_h),$$

where I_h is a standard interpolant into the finite element space $\mathcal{V}_h$. For brevity, instead of plotting the behavior of each $\delta_{i,h}^{(j)}$ for all i, j, we plot the behavior of their sum

$$d_h = \sum_{i=1}^{3} \sum_{j=1}^{2} \delta_{i,h}^{(j)}$$

for decreasing mesh sizes h and increasing polynomial degrees p in Figure 1. In the same figure panel, we also display the observed errors in the computed eigenvalues in Λ_h by plotting the Hausdorff distance $\mathrm{dist}(\Lambda, \Lambda_h)$ for various values of h and p.

Since $\delta_i^{(j)}$ should go to zero at the same rate as $\mathrm{gap}_{H_0^1(\Omega)}(E_h, E)$ and since the interpolation errors are of the same order as the gap, we expect d_h to go to zero as $h \to 0$ at the same rate as $\mathrm{gap}_{H_0^1(\Omega)}(E_h, E)$. From Figure 1a, we observe that d_h appears to converge to 0 at the rate $O(h^p)$ for $p = 1, 2$, and 3. Since the eigenfunctions on the unit square are analytic, Assumption 5 holds for this example with *any* $s_E > 0$. Therefore, our observation on the rate of convergence of d_h is in agreement with the gap estimate (31) of Theorem 3.6. Figure 1b shows that as h decreases, $\mathrm{dist}(\Lambda, \Lambda_h)$ decreases to 0 at the rate $O(h^{2p})$ for $p = 1, 2$, and 3. This is also in good agreement with the eigenvalue error estimate (32) of Theorem 3.6.

The results presented above using the DPG discretization are comparable to those found in [9] using the FEAST algorithm with the standard finite element discretization of comparable orders.

Remark 4.1. In other unreported experiments, we found that setting Y_h to

$$\tilde{Y}_h = \{y \in H^1(\Omega_h) : y|_K \in P_{p+1}(K)\}$$

also gave the same convergence rates. This indicates that the space dictated by the theory, namely $Y_h = \{y \in H^1(\Omega_h) : y|_K \in P_{p+3}(K)\}$, might be overly conservative. We already noted one approach to improve the estimates in Remark 3.3. Another approach might be through a perturbation argument, as the theory in [8] proves the error estimate of Lemma 3.2 at $z = 0$ even when Y_h is replaced by $\tilde{Y}_h$.

4.2. Convergence rates on an L-shaped domain.
In this example, we consider the Dirichlet eigenvalues of the L-shaped domain $\Omega = (0, 2) \times (0, 2) \setminus [1, 2] \times [1, 2]$ enclosed within a circular contour of radius $\gamma = 8$ centered at $y = 15$. The first three Dirichlet eigenvalues

are enclosed in this contour and we are interested in determining the eigenvalue error and numerical order of convergence for these. We use the results reported in [22] as our reference eigenvalues, namely $\lambda_1 \approx 9.6397238$, $\lambda_2 \approx 15.197252$, and $\lambda_3 = 2\pi^2$.

With the above values of λ_i (displayed up to the digits the authors of [22] claimed confidence in), we define $\text{ERR}(h) = |\lambda_{i,h} - \lambda_i|$, where $\lambda_{1,h} \leqslant \lambda_{2,h} \leqslant \lambda_{3,h}$ are the approximate eigenvalues obtained by FEAST. Then we define the numerical order of convergence (NOC) as $\text{NOC}(h) = \log(\text{ERR}(2h)/\text{ERR}(h))/\log(2)$.

We perform our convergence study, as in the unit square case, using a sequence of uniformly refined meshes, starting from a mesh size of $h = 2^{-2}$ and ending with a mesh size of $h = 2^{-6}$. In this example we confine the scope of our convergence study to polynomial degree $p = 2$. Further mesh refinements or higher degrees are not studied because the exact eigenvalues are only available to limited precision and errors below this precision cannot be used to surmise convergence rates accurately. The observations are compiled in Table 1.

From the first column of Table 1, we find that the first eigenvalue is observed to converge at a rate of approximately 4/3. For polygonal domains, its well known that Assumption 5 holds with any positive s less than the π/α where α is the largest of the interior angles at the vertices of the polygon. Clearly $\alpha = 3\pi/2$ for our L-shaped Ω. The eigenfunction corresponding to the first eigenvalue is known to be limited by this regularity, so s_E may be chosen to be any positive number less than 2/3. Therefore, the observed convergence rate of 4/3 for the first eigenvalue is in agreement with the rate of $2\min(p, s_E)$ established in Theorem 3.6. Although Theorem 3.6 does not yield improved convergence rates for the other eigenvalues, which have eigenfunctions of higher regularity, we observe from the remaining columns of Table 1 that in practice we do observe higher order convergence rates. E.g., the eigenfunction corresponding to $\lambda_3 = 2\pi^2$ is analytic and we observed that the corresponding eigenvalue converges at a rate $O(h^{2p})$ that is not limited by s_E.

5. Application to optical fibers

Double-clad step-index optical fibers have resulted in numerous technological innovations. Although originally intended to carry energy in a single mode, for increased power operation large mode area (LMA) fibers are now being sold extensively. LMA fibers usually have multiple guided modes. In this section, we show how to use the method we developed in the previous sections to compute such modes. We begin by showing that the problem of computing the fiber modes can be viewed as a problem of computing an eigenvalue cluster of an operator of the form discussed in Subsection 3.4.

These optical fibers have a cylindrical core of radius r_{core} and a cylindrical cladding region enveloping the core, extending to radius r_{clad}. We set up our axes so that the longitudinal direction of the fiber is the z-axis. The transverse coordinates will be denoted x, y while using Cartesian coordinates and the eigenvalue problem will be posed in these coordinates. Thus the space dimension (previously denoted by n) will be fixed to 2 in this section, so denoting the refractive index of the fiber by n in this section causes no confusion. We have in mind fibers whose refractive index $n(x, y)$ is a piecewise constant function, equalling n_{core} in the core, and n_{clad} in the cladding region ($n_{\text{clad}} < n_{\text{core}}$). The guided modes, also called the transverse core modes, decay exponentially in the cladding region.

These *modes* of the fiber, which we denote by $\varphi_l(x, y)$, are non-trivial functions that, together with their accompanying (positive) *propagation constants* β_l, solve

$$(40a) \qquad\qquad (\Delta + k^2 n^2)\varphi_l = \beta_l^2 \varphi_l, \qquad r < r_{\text{core}},$$

where k is a given wave number of the signal light, $\Delta = \partial_{xx} + \partial_{yy}$ denotes the Laplacian in the transverse coordinates x, y. Since the guided modes decay exponentially in the cladding, and since the cladding radius is typically many times larger than the core, we supplement (40a) with zero Dirichlet boundary conditions at the end of the cladding:

$$(40b) \qquad\qquad \varphi_l = 0, \qquad r = r_{\text{core}}.$$

Since the spectrum of the Dirichlet operator Δ lies in the negative real axis and has an accumulation point at $-\infty$, we expect to find only finitely many $\lambda_l \equiv \beta_l^2 > 0$ satisfying (40). This finite collection of eigenvalues λ_l form our eigenvalue cluster Λ in this application, and the corresponding eigenspace E is the span of the modes φ_l.

From the standard theory of step-index fibers [17], it follows that the propagation constants β_l of guided modes satisfy

$$n_{\text{clad}}^2 k^2 < \beta_l^2 < n_{\text{core}}^2 k^2.$$

Thus, having a pre-defined search interval, the computation of the eigenpairs (λ_l, φ_l) offers an example very well-suited for applying the FEAST algorithm. Moreover, since separation of variables can be employed to calculate the exact solution in terms of Bessel functions, we are able to perform convergence studies as well. Below, we apply the algorithm to a realistic fiber using the previously described DPG discretization of the resolvent of the Helmholtz operator $\Delta + k^2 n^2$ with Dirichlet boundary conditions to a realistic fiber.

The fiber we consider is the commercially available ytterbium-doped Nufern$^{\text{TM}}$ (nufern.com) fiber, whose typical parameters are

$$(41) \qquad n_{\text{core}} = 1.45097, \;\; n_{\text{clad}} = 1.44973, \;\; r_{\text{core}} = 0.0125 \text{ m}, \;\; r_{\text{clad}} = 16 r_{\text{core}}.$$

The typical operating wavelength for signals input to this fiber is 1064 nanometers, so we set the wavenumber to $k = (2\pi/1.064) \times 10^6$. Due to the small fiber radius, we compute after scaling the eigenproblem (40) to the unit disc $\hat{\Omega} = \{r < 1\}$, i.e., we compute modes $\hat{\varphi}_l : \hat{\Omega} \to \mathbb{C}$ satisfying $(\Delta + k^2 n^2 r_{\text{clad}}^2)\hat{\varphi}_l = r_{\text{clad}}^2 \beta_l^2 \hat{\varphi}_l$ in $\hat{\Omega}$ and $\hat{\varphi}_l = 0$ on $\partial\hat{\Omega}$. As in the previous section, all results here are generated using our code [10] built atop NGSolve [20]. Note that all experiments in this section are performed using the reduced $\tilde{Y}_h$ mentioned in Remark 4.1.

Results from the computation are given in Figures 2 and 3. Note that the elements whose boundary intersects the core or cladding boundary are isoparametrically curved to minimize boundary representation errors – see Figures 2a and 2b. The modes are localized near the core region, so the mesh is designed to be finer there. A six dimensional eigenspace was found. The computed basis for the 6-dimensional space of modes, obtained using polynomial degree $p = 6$, are shown (zoomed in near the core region) in the plots of Figure 3. The mode e_6 shown in Figure 3f is considered the "fundamental mode" for this fiber, also called the LP01 mode in the optics literature [17].

We also conducted a convergence study. We began with a mesh whose approximate mesh size in the core region is $h_c = 1/16$. We performed three uniform mesh refinements, where each refinement halved the mesh size. After each refinement, the elements intersecting the core or cladding boundary were curved again using the geometry information. Using the DPG discretization and $N = 16$ quadrature points for the contour integral, we computed the 6 eigenvalues, denoted by $\hat{\lambda}_l^h$, and compared them with the exact eigenvalues on the scaled domain, denoted by $\hat{\lambda}_l = r_{\text{core}}^2 \beta_l^2$. For the parameter values set in (41), there are six such $\hat{\lambda}_l$ (counting multiplicities) whose approximate values are $\hat{\lambda}_1 = 2932065.0334243$, $\hat{\lambda}_2 = \hat{\lambda}_3 =$
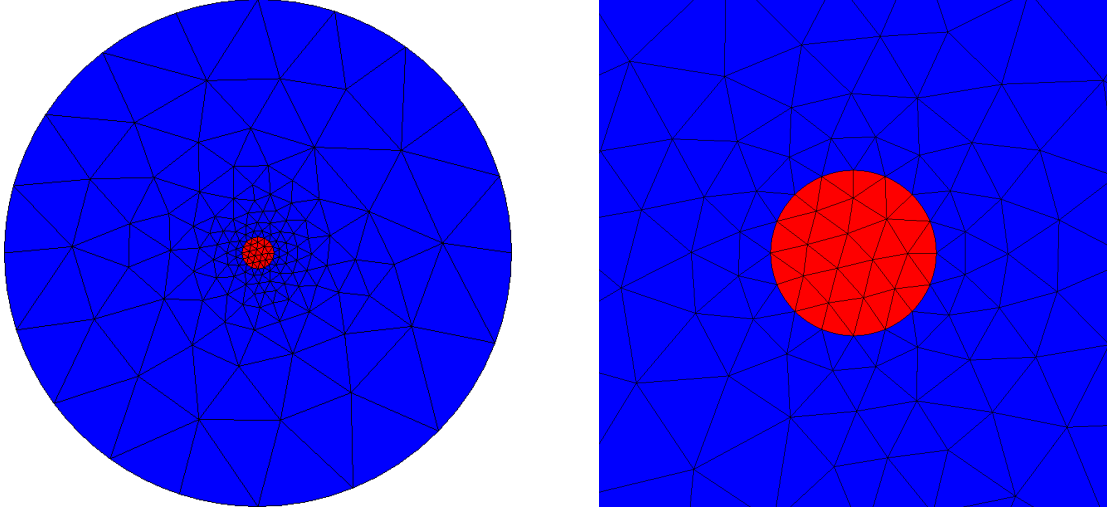
(A) The mesh with curved elements adjacent to the core and cladding boundaries.

(B) Zoomed-in view of the mesh in Figure 2a near the core.

FIGURE 2. The mesh used for computing modes of the ytterbium-doped fiber.

(A) φ_1^h

(B) φ_2^h

(C) φ_3^h

(D) φ_4^h
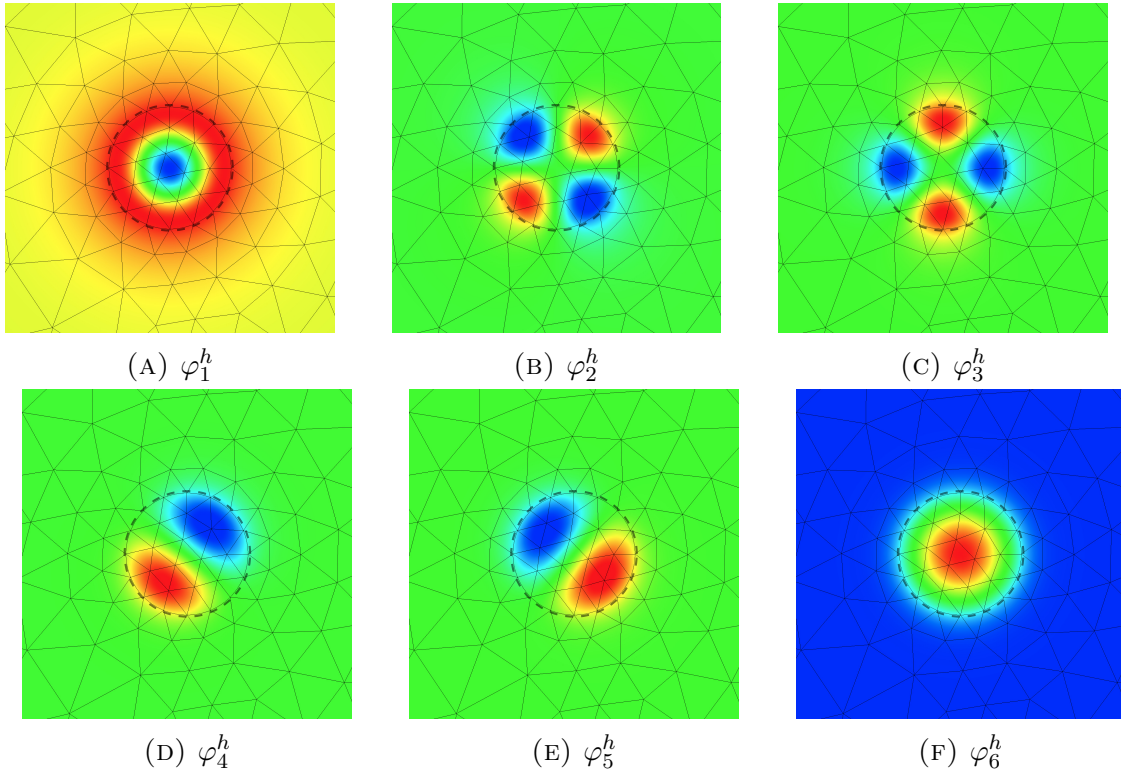
(E) φ_5^h

(F) φ_6^h

FIGURE 3. A close view of the approximate eigenfunctions φ_j^h computed by FEAST for the ytterbium-doped fiber. The boundary of the fiber core region is marked by dashed black circles.

core h	e_1	NOC	e_2	NOC	e_3	NOC	e_4	NOC	e_5	NOC	e_6	NOC
h_c	1.26e-07	–	2.01e-07	–	1.81e-07	–	4.99e-08	–	4.37e-08	–	1.72e-08	–
$h_c/2$	9.42e-09	3.7	1.63e-08	3.6	1.32e-08	3.8	6.46e-09	3.0	4.84e-09	3.2	3.38e-09	2.4
$h_c/4$	1.17e-10	6.3	2.13e-10	6.3	1.80e-10	6.2	7.03e-11	6.5	4.84e-11	6.6	3.64e-11	6.5
$h_c/8$	9.16e-14	10.3	1.33e-12	7.3	3.06e-13	9.2	3.75e-13	7.6	6.87e-13	6.1	6.69e-14	9.1

TABLE 2. Convergence rates of the fiber eigenvalues

2932475.1036310, $\hat{\lambda}_4 = \hat{\lambda}_5 = 2934248.1978369$, $\hat{\lambda}_6 = 2935689.8561775$. Fixing $p = 3$, we report the relative eigenvalue errors

$$e_l = \frac{|\hat{\lambda}_l - \hat{\lambda}_l^h|}{\hat{\lambda}_l^h}$$

in Table 2 for each l (columns) and each refinement level (rows). A column next to an e_l-column indicates the numerical order of convergence (computed as described in Section 4). The observed convergence rates are somewhat near the order of 6 expected from the previous theory. The match in the rates is not as close as in the results from the "textbook" benchmark examples of Section 4, presumably because mesh curving may have an influence on the pre-asymptotic behavior. Since the relative error values have quickly approached machine precision, further refinements were not performed.

References

[1] W.-J. BEYN, *An integral method for solving nonlinear eigenvalue problems*, Linear Algebra Appl., 436 (2012), pp. 3839–3863.

[2] T. BOUMA, J. GOPALAKRISHNAN, AND A. HARB, *Convergence rates of the DPG method with reduced test space degree*, Computers and Mathematics with Applications, 68 (2014), pp. 1550–1561.

[3] T. BÜHLER AND D. A. SALAMON, *Functional Analysis*, American Mathematical Society, 2018.

[4] C. CARSTENSEN, L. DEMKOWICZ, AND J. GOPALAKRISHNAN, *A posteriori error control for DPG methods*, SIAM J Numer. Anal., 52 (2014), pp. 1335–1353.

[5] C. CARSTENSEN, L. DEMKOWICZ, AND J. GOPALAKRISHNAN, *Breaking spaces and forms for the DPG method and applications including Maxwell equations*, Computers and Mathematics with Applications, 72 (2016), pp. 494–522.

[6] C. CARSTENSEN AND F. HELLWIG, *Optimal convergence rates for adaptive lowest-order discontinuous Petrov-Galerkin schemes*, SIAM J. Numer. Anal., 56 (2018), pp. 1091–1111.

[7] L. DEMKOWICZ AND J. GOPALAKRISHNAN, *A class of discontinuous Petrov-Galerkin methods. Part II: Optimal test functions*, Numerical Methods for Partial Differential Equations, 27 (2011), pp. 70–105.

[8] L. DEMKOWICZ AND J. GOPALAKRISHNAN, *A primal DPG method without a first-order reformulation*, Computers and Mathematics with Applications, 66 (2013), pp. 1058–1064.

[9] J. GOPALAKRISHNAN, L. GRUBIŠIĆ, AND J. OVALL, *Spectral discretization errors in filtered subspace iteration*, Preprint: arXiv:1709.06694, (2018).

[10] J. GOPALAKRISHNAN AND B. Q. PARKER, *Pythonic FEAST*. Software hosted at Bitbucket: `https://bitbucket.org/jayggg/pyeigfeast`.

[11] J. GOPALAKRISHNAN AND W. QIU, *An analysis of the practical DPG method*, Mathematics of Computation, 83 (2014), pp. 537–552.

[12] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, no. 24 in Monographs and Studies in Mathematics, Pitman Advanced Publishing Program, Marshfield, Massachusetts, 1985.

[13] S. GÜTTEL, E. POLIZZI, P. T. P. TANG, AND G. VIAUD, *Zolotarev quadrature rules and load balancing for the FEAST eigensolver*, SIAM J. Sci. Comput., 37 (2015), pp. A2100–A2122.

[14] A. HORNING AND A. TOWNSEND, *Feast for differential eigenvalue problems*, arXiv preprint 1901.04533, (2019).

[15] T. Kato, *Perturbation theory for linear operators*, Classics in Mathematics, Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.

[16] E. Polizzi, *A density matrix-based algorithm for solving eigenvalue problems*, Phys. Rev. B 79, 79 (2009), p. 115112.

[17] G. A. Reider, *Photonics: An introduction*, Springer, 2016.

[18] Y. Saad, *Analysis of subspace iteration for eigenvalue problems with evolving matrices*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 103–122.

[19] K. Schmüdgen, *Unbounded self-adjoint operators on Hilbert space*, vol. 265 of Graduate Texts in Mathematics, Springer, Dordrecht, 2012.

[20] J. Schöberl, *NGSolve.* http://ngsolve.org.

[21] T. Sakurai and H. Sugiura, *A projection method for generalized eigenvalue problems using numerical integration*, in Proceedings of the 6th Japan-China Joint Seminar on Numerical Mathematics (Tsukuba, 2002), vol. 159:1, 2003, pp. 119–128.

[22] L.N. Trefethen and T. Betcke, *Computed eigenmodes of planar regions*, in Recent advances in differential equations and mathematical physics, vol. 412 of Contemp. Math., Amer. Math. Soc., Providence, RI, 2006, pp. 297-314.

Portland State University, PO Box 751, Portland, OR 97207-0751, USA
E-mail address: gjay@pdx.edu

University of Zagreb, Bijenička 30, 10000 Zagreb, Croatia
E-mail address: luka.grubisic@math.hr

Portland State University, PO Box 751, Portland, OR 97207-0751, USA
E-mail address: jovall@pdx.edu

Portland State University, PO Box 751, Portland, OR 97207-0751, USA
E-mail address: bqp2@pdx.edu