SoftFEM: revisiting the spectral finite element approximation of elliptic operators

Quanling Deng, Alexandre Ern

SoftFEM: revisiting the spectral finite element approximation of elliptic operators

Quanling Deng* Alexandre Ern†

Abstract

We propose, analyze mathematically, and study numerically a novel approach for the finite element approximation of the spectrum of second-order elliptic operators. The main idea is to reduce the stiffness of the problem by subtracting to the standard stiffness bilinear form a least-squares penalty on the gradient jumps across the mesh interfaces. This penalty bilinear form is similar to the known technique used to stabilize finite element approximations in various contexts, but it brings here a negative contribution. Since it reduces the stiffness of the problem, the resulting approximation technique is called softFEM. The two key advantages of softFEM over the standard Galerkin FEM are to improve the approximation of the eigenvalues in the upper part of the discrete spectrum and to reduce the condition number of the stiffness matrix. We derive a sharp upper bound on the softness parameter weighting the stabilization bilinear form so as to maintain coercivity for the softFEM bilinear form. Then we prove that softFEM delivers the same optimal convergence rates as the standard Galerkin FEM approximation for the eigenvalues and the eigenvectors. We next compare the discrete eigenvalues obtained when using Galerkin FEM and softFEM. Finally, a detailed analysis of linear softFEM for the 1D Laplace eigenvalue problem delivers a sensible choice for the softness parameter. With this choice, the stiffness reduction ratio scales linearly with the polynomial degree. Various numerical experiments illustrate the benefits of using softFEM over Galerkin FEM.
Mathematics Subjects Classification: 65N30, 65N35, 35J05

Keywords finite element method (FEM), Laplacian, spectral approximation, eigenvalues, stiffness, gradient-jump penalty

1 Introduction

The optimal approximation of eigenvalues and eigenfunctions from second-order elliptic spectral problems by means of Galerkin finite element methods (FEM) is well-established. We refer the reader to the seminal contributions in Vainikko [32, 33], Bramble and Osborn [4], Strang and Fix [31], Osborn [28], Descloux et al. [19, 20], Babuška

*Department of Mathematics, University of Wisconsin–Madison, Madison, WI 53706, USA. E-mail addresses: quanling.deng@math.wisc.edu; qdeng12@gmail.com

†CERMICS, Ecole des Ponts, 77455 Marne la Vallée cedex 2, and INRIA Paris, 75589 Paris, France. E-mail address: alexandre.ern@enpc.fr

and Osborn [2], and to the more recent reviews in [3, 21]. The approximation of elliptic spectral problems has also been studied by means of mixed finite element methods [12, 27, 26], discontinuous Galerkin methods [1, 23], hybridizable discontinuous Galerkin methods [14, 24], hybrid high-order methods [10, 13], and virtual element methods [22]. All of these methods deliver optimally-convergent approximations, but since the eigenfunctions become more and more oscillatory in the upper part of the spectrum, the approximation is accurate only in the lower part of the spectrum. In contrast, isogeometric analysis [15] delivers a more accurate approximation in the upper part of the spectrum (see also [17, 11, 18] for some recent improvements on the subject).

The goal of this work is to improve on the Galerkin FEM spectral approximation so as to increase the accuracy in the upper part of the spectrum. This goal is achieved by reducing the stiffness of the discrete spectral problem. With this in mind, we refer to the newly-coined method as *softFEM*. The idea is to subtract to the standard stiffness bilinear form a least-squares penalty on the gradient jumps across the mesh interfaces. Thus, the softFEM bilinear form is defined as

$$\hat{a}(\cdot,\cdot) \coloneqq a(\cdot,\cdot) - \eta s(\cdot,\cdot), \tag{1.1}$$

where $a(\cdot,\cdot)$ is the standard Galerkin FEM stiffness bilinear form, η is the so-called *softness* parameter, and $s(\cdot,\cdot)$ is the bilinear form penalizing the gradient jumps across the mesh interfaces. The idea behind softFEM shares some common ground with isogeometric analysis where the basis functions have at least C^1-smoothness. In softFEM, the same basis functions are used as in Galerkin FEM so that the smoothness is only C^0. However, by considering the bilinear form $\hat{a}(\cdot,\cdot)$ instead of $a(\cdot,\cdot)$, one avoids that the eigenfunctions associated with the upper part of the spectrum store too much energy in the gradient jumps across the mesh interfaces. This change is not needed for the eigenfunctions associated with the lower part of the spectrum since such functions are able to represent well on the given mesh the exact smooth eigenfunctions. We notice that the bilinear form $s(\cdot,\cdot)$ has been considered for the purpose of stabilization (i.e., leading to a positive contribution and not to a negative one as in the present work) in various contexts related, in particular, to advection-dominated advection-diffusion equations and to the Stokes equations [7, 9, 8]. Incidentally, we mention that the term softFEM has been used recently in [29] in a completely different context related to heuristic optimization and soft computing for solid mechanics.

To give the reader a first view on the benefits of softFEM over Galerkin FEM, we present in Figure 1 the relative eigenvalue and eigenfunction errors for the 1D Laplace eigenvalue problem (with Dirichlet boundary conditions) using Galerkin FEM and softFEM, a uniform mesh composed of $N^h = 100$ elements, and a polynomial degree $p \in \{1,2,3\}$, so that the total number of discrete eigenpairs is $N_p^h \coloneqq pN^h - 1$. The benefit of using softFEM is evident when looking at the upper part of the spectrum. Another salient advantage of softFEM with respect to Galerkin FEM is that softFEM tempers the condition number of the stiffness matrix. This can have practically important consequences in the context of explicit time-marching schemes for time-dependent PDEs by reducing the CFL constraint on the time step. In many situations we observe that the

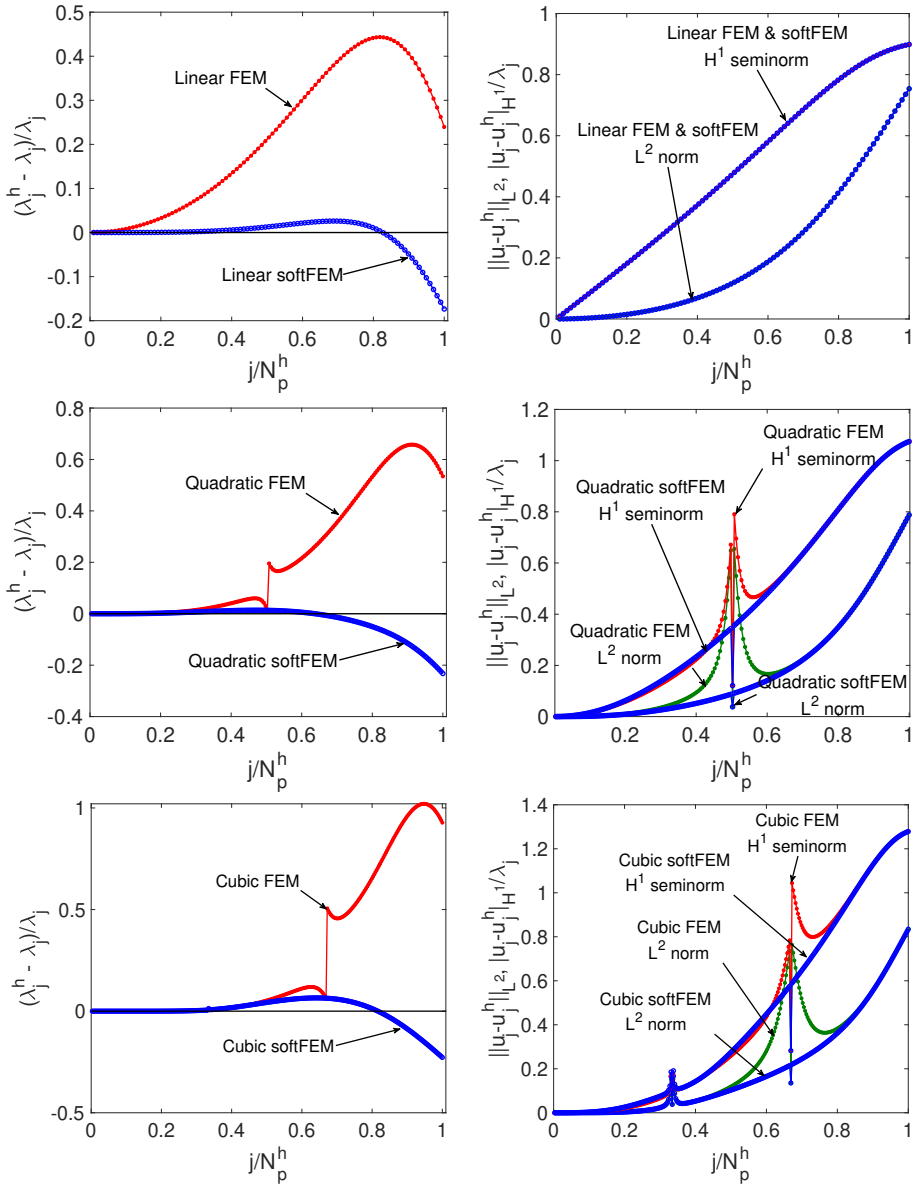

Figure 1: Relative eigenvalue (left) and eigenfunction (right) errors for the 1D Laplace eigenvalue problem when using Galerkin FEM and softFEM with $N^h = 100$ uniform elements and polynomial degrees $p \in \{1, 2, 3\}$. Upper row: $p = 1$; middle row: $p = 2$; bottom row: $p = 3$. The eigenfunction errors for linear Galerkin FEM and linear softFEM are the same as both discretization methods give the same eigenvectors (but not the same eigenvalues).

stiffness reduction ratio scales linearly with p and is of the order of $1 + \frac{p}{2}$.

The main mathematical results of this work can be summarized as follows. In

Theorem 2.2 we show that in order to maintain the coercivity of the softFEM bilinear form $\hat{a}(\cdot,\cdot)$, the softness parameter can be chosen so that $\eta \in [0, \eta_{\max})$, where the limit value depends on the polynomial degree p and the type of mesh (tensor-product or simplicial). Specifically $\eta_{\max} = \frac{1}{2p(p+1)}$ on tensor-product meshes and $\eta_{\max} = \frac{1}{2p(p+d-1)}$ on simplicial meshes (here $d \geq 2$ denotes the space dimension). This result is established by means of some discrete trace inequalities with sharp constants. In Theorem 2.3 we establish that softFEM maintains the same optimal convergence rates as Galerkin FEM. In Theorem 3.2 we prove for the 1D Laplace eigenvalue problem approximated by linear softFEM (i.e., $p = 1$), that the choice $\eta = \frac{1}{2(p+1)(p+2)} = \frac{1}{12}$ leads to superconvergence of the eigenvalue errors (quartic convergence rate instead of quadratic). We retain this choice for the value of the softness parameter in the rest of this work and notice that it is compatible with the maximum value $\eta_{\max}$ obtained in Theorem 2.2. Finally, in Theorem 2.5 we establish lower and upper bounds on the discrete softFEM eigenvalues by those approximated by Galerkin FEM. In particular, the lower bound shows that the optimal value for the stiffness reduction ratio should be $1 + \frac{p}{2}$ on tensor-product meshes and $1 + \frac{p}{4-d}$ on simplicial meshes with $d \in \{2, 3\}$. Both values are close to those observed in our numerical experiments.

The rest of this paper is organized as follows. Section 2 presents the exact spectral problem, its Galerkin FEM discretization, the softFEM approximation, as well as the following salient results concerning softFEM: coercivity (Theorem 2.2), error estimates (Theorem 2.3), and lower and upper bounds on the discrete eigenvalues (Theorem 2.5). Theorem 2.3 and Theorem 2.5 are proved in Section 2, but the proof of Theorem 2.2 is postponed to Section 5. Section 3 is concerned with the softFEM approximation of the 1D Laplace eigenvalue problem on uniform meshes. It contains the superconvergence result for softFEM (Theorem 3.2) motivating the choice $\eta = \frac{1}{2(p+1)(p+2)}$ for the softness parameter, and numerical experiments for various polynomial degrees illustrating the benefits of using softFEM with respect to Galerkin FEM both for the accuracy of the upper part of the spectrum and for the stiffness reduction. Section 4 collects more challenging numerical examples (Laplace eigenvalue problem in multiple dimensions, elliptic eigenvalue problem and non-uniform meshes for the 1D Laplace eigenvalue problem, and the use of simplicial meshes on the unit square and the L-shaped domain still for the Laplace eigenvalue problem) which corroborate the positive conclusions drawn on softFEM in Section 3. In Section 5, we first study discrete trace inequalities with sharp constants and then use these inequalities to prove Theorem 2.2. Concluding remarks are presented in Section 6.

2 Main idea and results

In this section we state the elliptic eigenvalue problem and describe its approximation by means of Galerkin FEM and softFEM. We then state the main results concerning softFEM.

2.1 Problem statement

Let Ω be a bounded, open subset of $\mathbb{R}^d$, $d \geq 1$, with Lipschitz boundary $\partial\Omega$. For simplicity, we assume in what follows that Ω is a polyhedron. We use standard notation for the Lebesgue and Sobolev spaces. For any measurable subset $S \subseteq \Omega$, we denote the L^2-inner product and norm as $(\cdot,\cdot)_S$ and $\|\cdot\|_S$, respectively, and the same notation is used for vector-valued fields. For any integer $m \geq 1$, we denote the H^m-norm and H^m-seminorm as $\|\cdot\|_{H^m(S)}$ and $|\cdot|_{H^m(S)}$, respectively.

We consider the following second-order elliptic eigenvalue problem with homogeneous Dirichlet boundary conditions: Find an eigenpair $(\lambda, u) \in \mathbb{R}^+ \times H^1_0(\Omega)$ such that $\|u\|_\Omega = 1$ and

$$\begin{aligned} -\nabla\cdot(\kappa\nabla u) &= \lambda u \quad \text{in } \Omega,\\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

with the diffusion coefficient $\kappa \in L^\infty(\Omega)$ uniformly bounded from below away from zero, and we set $\kappa_{\min} := \operatorname{ess\,inf}_{x\in\Omega}\kappa(x) > 0$. For $\kappa = 1$, the problem (2.1) reduces to the Laplace (Dirichlet) eigenvalue problem. The variational formulation of (2.1) is

$$a(u,w) = \lambda b(u,w), \quad \forall w \in H^1_0(\Omega), \tag{2.2}$$

with the bilinear forms

$$a(v,w) := (\kappa\nabla v, \nabla w)_\Omega, \qquad b(v,w) := (v,w)_\Omega. \tag{2.3}$$

The eigenvalue problem (2.1) has a countable set of eigenvalues $\lambda_j \in \mathbb{R}^+$ (see, for example, [5, Sec. 9.8])

$$0 < \lambda_1 < \lambda_2 \leq \lambda_3 \leq \dots$$

and an associated set of L^2-orthonormal eigenfunctions u_j, that is, $(u_j, u_k) = \delta_{jk}$, where δ_{jk} is the Kronecker delta. With (2.2) in mind, the normalized eigenfunctions are also orthogonal in the energy inner product since we have $a(u_j,u_k) = \lambda_j b(u_j,u_k) = \lambda_j\delta_{jk}$. In what follows we always sort the eigenvalues in ascending order counted with their order of algebraic multiplicity.

2.2 Galerkin FEM

Let $(\mathcal{T}_h)_{h>0}$ be a shape-regular sequence of meshes of Ω. A generic mesh element is denoted τ, its diameter h_τ, and its outward unit normal $\boldsymbol{n}_\tau$. We set $h := \max_{\tau\in\mathcal{T}_h} h_\tau$. To stay general, we consider both tensor-product meshes where the mesh elements are cuboids (and so is the domain Ω), and simplicial meshes where the mesh elements are simplices (triangles if $d = 2$, tetrahedra if $d = 3$). Let $p \geq 1$ be the polynomial degree. Let $\mathbb{P}_p(\tau)$ (resp., $\mathbb{Q}_p(\tau)$) be the space composed of the restriction to τ of polynomials of total degree at most p (resp., of degree at most p in each variable). For tensor-product meshes, the Galerkin finite element approximation space is defined as

$$V_p^h := \{v_h \in C^0(\overline{\Omega}) : v_h|_{\partial\Omega} = 0, \forall\tau\in\mathcal{T}_h, v_h|_\tau \in \mathbb{Q}_p(\tau)\}, \tag{2.4}$$

whereas for simplicial meshes, it is defined as

$$V_p^h := \{v_h \in C^0(\overline{\Omega}) : v_h|_{\partial\Omega} = 0, \forall \tau \in \mathcal{T}_h, v_h|_\tau \in \mathbb{P}_p(\tau)\}. \tag{2.5}$$

It is well-known that in both cases $V_p^h \subset H_0^1(\Omega)$.

The Galerkin FEM approximation of (2.1) seeks $(\lambda^h, u^h) \in \mathbb{R}^+ \times V_p^h$ such that $\|u^h\|_\Omega = 1$ and

$$a(u^h, w^h) = \lambda^h b(u^h, w^h), \quad \forall w^h \in V_p^h. \tag{2.6}$$

The algebraic realization of (2.6) follows by choosing basis functions $\{\phi_j^h\}_{j\in\{1,\dots,N_p^h\}}$ of V_p^h with $N_p^h := \dim(V_p^h)$ (typically, one considers nodal basis functions). This leads to the following generalized matrix eigenvalue problem (GMEVP):

$$\mathbf{K}\mathbf{U} = \lambda^h \mathbf{M}\mathbf{U}, \tag{2.7}$$

where $\mathbf{K}_{kl} := a(\phi_k^h, \phi_l^h)$ and $\mathbf{M}_{kl} := b(\phi_k^h, \phi_l^h)$, for all $k, l \in \{1, \dots, N_p^h\}$, are the entries of the stiffness and mass matrices, respectively, and $\mathbf{U} \in \mathbb{R}^{N_p^h}$ is the eigenvector collecting the components of u^h in the chosen basis.

2.3 SoftFEM

Since softFEM is defined by subtracting to the bilinear form $a(\cdot,\cdot)$ a least-squares penalty on the jumps of the normal derivatives across the mesh interfaces, we first introduce some useful notation. For all $\tau \in \mathcal{T}_h$, we define h_τ^0 to be the length of the smallest edge of τ if τ is a cuboid, whereas we set $h_\tau^0 := \frac{d|\tau|}{|\partial\tau|}$ if τ is a simplex. Let $\mathcal{F}_h^i$ be the collection of the mesh interfaces. For all $F \in \mathcal{F}_h^i$, we have $F = \partial\tau_1 \cap \partial\tau_2$ for two distinct mesh elements $\tau_1, \tau_2 \in \mathcal{T}_h$. We then set

$$h_F := \min(h_{\tau_1}^0, h_{\tau_2}^0), \qquad \kappa_F := \min(\kappa_{\tau_1}, \kappa_{\tau_2}), \tag{2.8}$$

with $\kappa_\tau := \operatorname{ess\,inf}_{x\in\tau}\kappa(x)$ (i.e., κ_F is the smallest value of κ on the two elements that share the interface F). Moreover, for any function $v^h \in V_p^h$, we define the jump of its normal derivative across F as

$$[\![\nabla v^h \cdot \boldsymbol{n}]\!]_F := \nabla v^h|_{\tau_1} \cdot \boldsymbol{n}_{\tau_1} + \nabla v^h|_{\tau_2} \cdot \boldsymbol{n}_{\tau_2}. \tag{2.9}$$

We drop the subscript F when the context is unambiguous.

The softFEM approximation of (2.1) seeks $(\hat{\lambda}^h, \hat{u}^h) \in \mathbb{R}^+ \times V_p^h$ such that $\|\hat{u}^h\|_\Omega = 1$ and

$$\hat{a}(\hat{u}^h, w^h) = \hat{\lambda}^h b(\hat{u}^h, w^h), \quad \forall w^h \in V_p^h, \tag{2.10}$$

where for all $v^h, w^h \in V_p^h$,

$$\hat{a}(\cdot,\cdot) := a(\cdot,\cdot) - \eta s(\cdot,\cdot) \qquad \text{with} \qquad s(v^h, w^h) := \sum_{F\in\mathcal{F}_h^i} \kappa_F h_F ([\![\nabla v^h \cdot \boldsymbol{n}]\!], [\![\nabla w^h \cdot \boldsymbol{n}]\!])_F, \tag{2.11}$$

and $\eta \geq 0$ is a parameter to be specified below. The terminology *softFEM* is motivated by the fact that the term $-\eta s(\cdot,\cdot)$ reduces the stiffness of the system. We refer to η as the *softness parameter*. We will see below that one can take $\eta \in [0, \eta_{\max})$ for some $\eta_{\max}$ depending on the polynomial degree p and the type of mesh elements so that the bilinear form $\hat{a}(\cdot,\cdot)$ remains coercive. When $\eta = 0$, softFEM reduces to FEM.

Similarly to Galerkin FEM, the algebraic realization of the softFEM approximation (2.10) leads to the GMEVP

$$\hat{\mathbf{K}}\hat{\mathbf{U}} = \hat{\lambda}^h \mathbf{M}\hat{\mathbf{U}}, \tag{2.12}$$

where $\hat{\mathbf{K}} := \mathbf{K} - \eta\mathbf{S}$ with $\mathbf{S}_{kl} := s(\phi_k^h, \phi_l^h)$, $\mathbf{K}$ and $\mathbf{M}$ are respectively the stiffness and mass matrices as in (2.7), and $\hat{\mathbf{U}}$ is the eigenvector collecting the components of $\hat{u}^h$ in the chosen basis $\{\phi_j^h\}_{j\in\{1,\dots,N_p^h\}}$ of V_p^h.

Remark 2.1 (Variants)**.** For $p \geq 2$, the stiffness can be further reduced by imposing least-squares penalties on higher-order derivative jumps. However, these additional terms increase the computational costs while our numerical experiments (not shown for brevity) indicate only a further marginal improvement in terms of spectral errors. We also mention the recent work [18] which penalizes both the higher-order derivatives as well as the mass bilinear form near the boundary to eliminate the so-called outliers in isogeometric spectral approximations.

2.4 Main results on softFEM

In this section we present our main results on softFEM. We first derive an upper bound on the softness parameter to ensure coercivity of the bilinear form $\hat{a}(\cdot,\cdot)$. To improve readability, the proof is postponed to Section 5.

Theorem 2.2 (Coercivity)**.** *Let $\hat{a}(\cdot,\cdot)$ be defined in* (2.11)*. Set $\eta_{\max} := \frac{1}{2p(p+1)}$ for tensor-product meshes with $d \geq 1$ and $\eta_{\max} := \frac{1}{2p(p+d-1)}$ for simplicial meshes with $d \geq 2$. Assume that the softness parameter $\eta \in [0, \eta_{\max})$. The following holds:*

$$\beta_1 |w^h|^2_{H^1(\Omega)} \leq \hat{a}(w^h, w^h), \qquad \forall w^h \in V_p^h, \tag{2.13}$$

with $\beta_1 := \kappa_{\min}(1 - \frac{\eta}{\eta_{\max}}) > 0$.

Let us now consider the convergence of eigenvalues and eigenfunctions for softFEM. The solution operator $T : L^2(\Omega) \to L^2(\Omega)$ associated with the elliptic eigenvalue problem (2.2) is such that for all $\phi \in L^2(\Omega)$, $T(\phi) \in H_0^1(\Omega) \subset L^2(\Omega)$ is uniquely defined by requiring that $a(T(\phi), w) = b(\phi, w)$ for all $w \in H_0^1(\Omega)$. Notice that T is selfadjoint and compact, and the elliptic regularity theory implies that there is $s \in (\frac{1}{2}, 1]$ such that T maps boundedly from $L^2(\Omega)$ into $H^{1+s}(\Omega)$. Moreover (λ, u) is an eigenpair of (2.2) if and only if (μ, u) is an eigenpair of T with $\mu = \lambda^{-1}$.

Theorem 2.3 (Eigenvalue and eigenfunction errors)**.** *Let $(\lambda_j, u_j) \in \mathbb{R}^+ \times H_0^1(\Omega)$ solve* (2.2) *and let $(\hat{\lambda}_j^h, \hat{u}_j^h) \in \mathbb{R}^+ \times V_p^h$ solve* (2.10) *with the normalizations $\|u_j\|_\Omega = 1$ and $\|\hat{u}_j^h\|_\Omega = 1$. Let $s \in (\frac{1}{2}, 1]$ be the index of elliptic regularity. Assume that there is $t \in [s, p]$*

and a constant C_t such that one has the following smoothness property: $\|\phi\|_{H^{1+t}(\Omega)} + \|T(\phi)\|_{H^{1+t}(\Omega)} \le C_t\|\phi\|_\Omega$ for all $\phi \in G_j := ker(\mu_j I - T)$ with $\mu_j := \lambda_j^{-1}$. Then, the following holds:

$$\left|\hat{\lambda}_j^h - \lambda_j\right| \le Ch^{2t}, \qquad |u_j - \hat{u}_j^h|_{H^1(\Omega)} \le Ch^t, \tag{2.14}$$

where C is a positive constant independent of the mesh-size h. The convergence rates are optimal whenever $t = p$.

Proof. We cannot apply directly the classical theory for error analysis derived in [2, Thm. 7.2 & 7.4] since the softFEM bilinear form $\hat{a}(\cdot,\cdot)$ differs from $a(\cdot,\cdot)$. Instead, we can apply the extension of this theory presented in [21, Chap. 48] to finite element approximations with so-called variational crimes. We can work on the extended space $Y^h := V_p^h + H^{1+s}(\Omega)$ and establish the boundedness of $\hat{a}$ on $Y^h \times Y^h$ using the H^1-seminorm augmented by $s(\cdot,\cdot)^{\frac{1}{2}}$. Optimal approximation properties in this norm are readily derived for smooth functions. Moreover, consistency holds true since we have $s(u_j, y) = 0$ for all $y \in Y^h$ because $s > \frac{1}{2}$. This implies the above error estimates. □

Remark 2.4 (Pythagorean identity)**.** A classical identity relating the eigenvalue and eigenfunction errors (see, e.g., [31, Chap. 6]) is

$$\|u_j - \hat{u}_j^h\|_E^2 = \lambda_j\|u_j - \hat{u}_j^h\|_\Omega^2 + \hat{\lambda}_j^h - \lambda_j,$$

where $\|\cdot\|_E^2 := \hat{a}(\cdot,\cdot) \ge \beta_1|\cdot|_{H^1(\Omega)}^2$ owing to Lemma 2.2.

Our third main result quantifies the stiffness reduction by softFEM for one particular choice of the softness parameter η that is further motivated in Section 3 (see, in particular, Theorem 3.2), namely $\eta = \frac{1}{2(p+1)(p+2)}$. Notice that $\eta < \eta_{\max} = \frac{1}{2p(p+1)}$ for tensor-product meshes and that $\eta < \eta_{\max} = \frac{1}{2p(p+d-1)}$ with $d \in \{2,3\}$ on simplicial meshes.

Theorem 2.5 (Eigenvalue lower and upper bounds)**.** *Assume that $\eta = \frac{1}{2(p+1)(p+2)}$. Assume that $d \in \{2,3\}$ if simplicial meshes are used. Let $j \in \mathbb{N}$, let $(\lambda_j^h, u_j^h) \in \mathbb{R} \times V_p^h$ solve (2.6), and let $(\hat{\lambda}_j^h, \hat{u}_j^h) \in \mathbb{R} \times V_p^h$ solve (2.10) with the normalizations $\|u_j^h\|_\Omega = 1$ and $\|\hat{u}_j^h\|_\Omega = 1$. The following holds:*

$$\gamma_p\lambda_j^h \le \hat{\lambda}_j^h < \lambda_j^h, \tag{2.15}$$

with $\gamma_p := \frac{2}{p+2}$ on tensor-product meshes and $\gamma_p := \frac{4-d}{p+4-d}$ on simplicial meshes.

Proof. For all $v^h \in V_p^h \setminus \{0\}$, let us define the Rayleigh quotients

$$R(v^h) := \frac{a(v^h, v^h)}{b(v^h, v^h)}, \qquad \hat{R}(v^h) := \frac{\hat{a}(v^h, v^h)}{b(v^h, v^h)}$$

As shown in Section 5 (see (5.8)), we have

$$(1 - 2p(p+1)\eta)a(v^h, v^h) \le \hat{a}(v^h, v^h) < a(v^h, v^h)$$

on tensor-product meshes and

$$(1-2p(p+d-1)\eta)a(v^h,v^h) \le \hat{a}(v^h,v^h) < a(v^h,v^h)$$

on simplicial meshes. With the choice $\eta = \frac{1}{2(p+1)(p+2)}$, a direct calculation shows that

$$\gamma_p a(v^h,v^h) \le \hat{a}(v^h,v^h) < a(v^h,v^h),$$

with γ_p defined in the assertion, which readily implies that

$$\gamma_p R(v^h) \le \hat{R}(v^h) < R(v^h). \tag{2.16}$$

Let V_j denote the set of the subspaces of V_p^h of dimension $j \ge 1$. Then classical results on the Rayleigh quotient imply that

$$\lambda_j^h = \min_{E_j \in V_j} \max_{v^h \in E_j} R(v^h), \qquad \hat{\lambda}_j^h = \min_{E_j \in V_j} \max_{v^h \in E_j} \hat{R}(v^h).$$

The bounds in (2.15) then readily follow from (2.16). □

Since the stiffness matrices $\mathbf{K}$ and $\hat{\mathbf{K}}$ are symmetric, their condition numbers are given by

$$\sigma := \frac{\lambda_{\max}^h}{\lambda_{\min}^h}, \qquad \hat{\sigma} := \frac{\hat{\lambda}_{\max}^h}{\hat{\lambda}_{\min}^h}, \tag{2.17}$$

where $\lambda_{\max}^h, \hat{\lambda}_{\max}^h$ are the largest eigenvalues and $\lambda_{\min}^h, \hat{\lambda}_{\min}^h$ are the smallest eigenvalue of the matrix eigenvalue problems (2.7) and (2.12) that are associated with (2.6) and (2.10), respectively. We define the *stiffness reduction ratio* of softFEM with respect to Galerkin FEM as

$$\rho := \frac{\sigma}{\hat{\sigma}} = \frac{\lambda_{\max}^h}{\hat{\lambda}_{\max}^h} \cdot \frac{\hat{\lambda}_{\min}^h}{\lambda_{\min}^h}. \tag{2.18}$$

In general, for Galerkin FEM and softFEM with sufficient elements (i.e., as $h \to 0$), one has $\lambda_{\min}^h \approx \hat{\lambda}_{\min}^h$. Thus, the stiffness reduction ratio depends only on the largest eigenvalues for both methods. Since softFEM leads to a smaller largest eigenvalue, softFEM lowers the condition number of the stiffness matrix, i.e., $\rho \ge 1$. We define the *asymptotic stiffness reduction ratio* of softFEM with respect to Galerkin FEM as

$$\rho_\infty := \lim_{h \to 0} \frac{\lambda_{\max}^h}{\hat{\lambda}_{\max}^h}. \tag{2.19}$$

Theorem 2.5 shows that for $\eta = \frac{1}{2(p+1)(p+2)}$, the best possible asymptotic stiffness reduction ratio is $1 + \frac{p}{2}$ on tensor-product meshes and $1 + \frac{p}{4-d}$ on simplicial meshes with $d \in \{2,3\}$. Notice that for both types of meshes, this value grows linearly with p. Our numerical experiments reported in Section 3.2 for the 1D Laplace eigenvalue problem show that the asymptotic stiffness reduction ratio is indeed $\rho_\infty = 1 + \frac{p}{2}$. Moreover, the values of the asymptotic stiffness reduction ratio observed in the more general situations

studied in Section 4 are also close to the predictions of Theorem 2.5. Finally, we define the *stiffness reduction percentage* of softFEM with respect to Galerkin FEM as

$$\varrho = 100\frac{\sigma - \hat{\sigma}}{\sigma}\,\% = 100(1-\rho^{-1})\,\%, \tag{2.20}$$

and the *asymptotic stiffness reduction percentage* as $\varrho_\infty := 100(1-\rho_\infty^{-1})\,\%$, respectively.

Remark 2.6 (SoftFEM eigenvalues)**.** It is well-known that for Galerkin FEM, one has $\lambda_j \le \lambda_j^h$ for all $j \ge 1$, but this is not necessarily the case for softFEM. Our numerical experiments indicate that softFEM approximates the exact eigenvalues from above in the low-frequency region and from below in the high-frequency region.

3 Laplace eigenvalue problem in 1D

In this section, we focus on the spectral problem (2.1) with $\Omega := (0,1)$ and $\kappa := 1$, that is, on the 1D Laplace eigenvalue problem. In this case, the problem (2.1) has exact eigenvalues and L^2-normalized eigenfunctions

$$\lambda_j = j^2\pi^2 \quad \text{and} \quad u_j(x) = \sqrt{2}\sin(j\pi x), \quad j = 1,2,\dots, \tag{3.1}$$

respectively. We partition the interval $\Omega = (0,1)$ into N^h uniform elements so that the mesh size $h = 1/N^h$. We first focus on the case of linear finite elements ($p = 1$) and derive some analytical results showing that in this case the optimal choice for the softness parameter is $\eta = \frac{1}{12}$, that is, $\eta = \frac{1}{2(p+1)(p+2)}$ for $p = 1$. Then we present numerical experiments for this choice of the softness parameter and various polynomial degrees.

3.1 Analytical results for linear softFEM

The advantage of using linear elements is that it is possible to compute analytically the eigenvalues and eigenvectors for Galerkin FEM and softFEM. Firstly, it is well-known that the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ with $p = 1$ lead to the following stiffness and mass matrices:

$$\mathbf{K} = \frac{1}{h}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}, \qquad \mathbf{M} = h\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & & & \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ & & & \frac{1}{6} & \frac{2}{3} \end{bmatrix}, \tag{3.2}$$

which are of order $(N^h-1)\times(N^h-1)$. The bilinear form $s(\cdot,\cdot)$ leads to the matrix

$$\mathbf{S} = \frac{1}{h}\begin{bmatrix} 5 & -4 & 1 & & & & \\ -4 & 6 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 6 & -4 \\ & & & & 1 & -4 & 5 \end{bmatrix}, \tag{3.3}$$

which is also of order $(N^h-1)\times(N^h-1)$. Recall that we then have $\hat{\mathbf{K}} := \mathbf{K} - \eta\mathbf{S}$, and that according to Theorem 2.2, we must take the softness parameter $\eta \in [0, \eta_{\max})$ with $\eta_{\max} = \frac{1}{2p(p+1)} = \frac{1}{6}$ since $p = 1$ here.

Lemma 3.1 (Analytical eigenvalues and eigenvectors)**.** (i) *Galerkin FEM approximation: The GMEVP* $\mathbf{KU} = \lambda^h\mathbf{MU}$ *has eigenpairs* $(\lambda_j^h, \mathbf{U}_j)$ *for all* $j \in \{1,\dots,N^h-1\}$ *with*

$$\lambda_j^h = \frac{6}{h^2}\frac{1-\cos(t_j)}{2+\cos(t_j)}, \qquad \mathbf{U}_j = c_j\big(\sin(kt_j)\big)_{k\in\{1,\dots,N^h-1\}}, \tag{3.4}$$

with $t_j := j\pi h$ *and some normalization constant* $c_j > 0$. (ii) *SoftFEM approximation: The GMEVP* $\hat{\mathbf{K}}\hat{\mathbf{U}} = \hat{\lambda}^h\mathbf{M}\hat{\mathbf{U}}$ *has eigenpairs* $(\hat{\lambda}_j^h, \hat{\mathbf{U}}_j)$ *for all* $j \in \{1,\dots,N^h-1\}$ *with*

$$\hat{\lambda}_j^h = \frac{6}{h^2}\frac{1+3\eta-(1+4\eta)\cos(t_j)+\eta\cos(2t_j)}{2+\cos(t_j)}, \qquad \hat{\mathbf{U}}_j = \mathbf{U}_j. \tag{3.5}$$

Proof. The result (3.4) is well-known; see, for example, [3, Sec. 2] or [25, Sec. 4], whereas the result (3.5) follows for instance from an application of [16, Thm. 2.1]. □

An interesting consequence of (3.4)-(3.5) is that for linear softFEM, the stiffness reduction ratio and the asymptotic stiffness reduction ratio are

$$\rho = \frac{\lambda_{\max}^h}{\hat{\lambda}_{\max}^h}\cdot\frac{\hat{\lambda}_{\min}^h}{\lambda_{\min}^h} = \frac{5+\cos(\pi h)}{5-\cos(\pi h)}, \qquad \rho_\infty = \lim_{h\to 0}\frac{5+\cos(\pi h)}{5-\cos(\pi h)} = \frac{3}{2}. \tag{3.6}$$

Thus, asymptotically, linear softFEM reduces the stiffness of Galerkin FEM by about 33.3%.

For all $\eta \in [0, \eta_{\max})$ with $\eta_{\max} = \frac{1}{2p(p+1)} = \frac{1}{6}$, Theorem 2.3 shows that one should expect a quadratic convergence rate for the discrete eigenvalues. We now show that for the specific choice $\eta = \frac{1}{2(p+1)(p+2)} = \frac{1}{12}$, one obtains a quartic convergence rate, uniformly for all the discrete eigenvalues.

Theorem 3.2 (Eigenvalue superconvergence)**.** *Let* λ_j *be the* j*-th exact eigenvalue of* (2.1) *and let* $\hat{\lambda}_j^h$ *be the* j*-th approximate eigenvalue using linear softFEM. Assume that* $\eta = \frac{1}{2(p+1)(p+2)} = \frac{1}{12}$. *The following holds:*

$$\frac{|\hat{\lambda}_j^h - \lambda_j|}{\lambda_j} < \frac{1}{360}(j\pi h)^4, \qquad \forall j \in \{1,\dots,N^h-1\}. \tag{3.7}$$

Proof. The exact eigenvalues λ_j are given in (3.1), and the approximate eigenvalues $\hat{\lambda}_j^h$ are given in (3.5). To motivate the result of Theorem 3.2, we observe that applying a Taylor expansion to $\hat{\lambda}_j^h$, we obtain (recall that $t_j := j\pi h$)

$$\frac{\hat{\lambda}_j^h - \lambda_j}{\lambda_j} = \frac{1-12\eta}{12} t_j^2 + \frac{1}{360} t_j^4 - \frac{17-84\eta}{60480} t_j^6 + \mathcal{O}(t_j^8),$$

showing that the choice $\eta = \frac{1}{12}$ leads to a cancellation of the dominant term in the expansion and that the sixth-order term has a negative coefficient. More rigorously, using (3.1), (3.5), and algebraic manipulations, we infer that

$$\frac{|\hat{\lambda}_j^h - \lambda_j|}{\lambda_j} = \left| \frac{9 - (2+\cos(t_j))^2}{t_j^2(2+\cos(t_j))} - 1 \right|.$$

Since t_j samples the interval $(0,\pi)$, we can consider a continuous variable $t \in (0,\pi)$ and prove more generally that

$$\left| \frac{9 - (2+\cos(t))^2}{t^2(2+\cos(t))} - 1 \right| < \frac{1}{360} t^4,$$

or, equivalently, that

$$-t^6(2+\cos(t)) < 3240 - 360(2+\cos(t))^2 - 360t^2(2+\cos(t)) < t^6(2+\cos(t)),$$

for all $t \in (0,\pi)$. For the first inequality, we notice that the function

$$f(t) := t^6(2+\cos(t)) + 3240 - 360(2+\cos(t))^2 - 360t^2(2+\cos(t))$$

is increasing on $(0,t_0)$ and decreasing on (t_0,π) with $t_0 \approx 2.79911$, that $f(0) = 0$ and $f(\pi) \approx 0.800921$. The minimum value of f in $(0,\pi)$ is thus $f(0) = 0$. For the second inequality, we notice that the function

$$g(t) := t^6(2+\cos(t)) - 3240 + 360(2+\cos(t))^2 + 360t^2(2+\cos(t))$$

is increasing on $(0,\pi)$ and that $g(0) = 0$. This completes the proof. □

3.2 Numerical results for arbitrary-order softFEM in 1D

In this section we explore numerically softFEM for various polynomial degrees $p \geq 1$ using in all cases the softness parameter $\eta = \frac{1}{2(p+1)(p+2)}$.

Recall that Figure 1 shows the relative eigenvalue and eigenfunction errors for Galerkin FEM and softFEM with $N^h = 100$ uniform elements and polynomial orders $p \in \{1,2,3\}$. Notice that there are $N_p^h := pN^h - 1$ eigenpairs both for Galerkin FEM and for softFEM. We refer the reader to Section 3.3 for a brief discussion on the structure of the discrete spectrum for Galerkin FEM, including the notions of acoustic/optical branches and stopping bands. The improvement offered by softFEM over Galerkin FEM

p	N^h	$\frac{\lvert\hat{\lambda}_1^h-\lambda_1\rvert}{\lambda_1}$	$\lvert u_1-\hat{u}_1^h\rvert_{H^1}$	$\lVert u_1-\hat{u}_1^h\rVert_{L^2}$	$\frac{\lvert\hat{\lambda}_6^h-\lambda_6\rvert}{\lambda_6}$	$\lvert u_6-\hat{u}_6^h\rvert_{H^1}$	$\lVert u_6-\hat{u}_6^h\rVert_{L^2}$
1	8	6.54e-5	3.58e-1	5.85e-3	2.10e-2	1.40e1	3.56e-1
	16	4.12e-6	1.78e-1	1.44e-3	4.80e-3	6.63	6.06e-2
	32	2.58e-7	8.91e-2	3.60e-4	3.27e-4	3.23	1.35e-2
	64	1.61e-8	4.45e-2	8.98e-5	2.08e-5	1.61	3.27e-3
	rate	4.00	1.00	2.01	3.38	1.04	2.25
2	4	4.38e-4	7.57e-2	2.54e-3	3.08e-2	1.37e1	2.82e-1
	8	3.15e-5	1.84e-2	3.40e-4	1.11e-2	3.95	4.47e-2
	16	2.04e-6	4.53e-3	4.33e-5	1.80e-3	1.04	7.78e-3
	32	1.29e-7	1.13e-3	5.43e-6	1.50e-4	2.52e-1	1.11e-3
	64	8.06e-9	2.82e-4	6.80e-7	1.02e-5	6.15e-2	1.45e-4
	rate	3.94	2.02	2.97	2.93	1.96	2.72
3	4	1.16e-7	5.82e-3	8.08e-5	4.32e-2	5.24	1.04e-1
	8	4.47e-10	7.19e-4	4.80e-6	7.64e-4	9.12e-1	9.29e-3
	16	2.08e-12	8.96e-5	2.96e-7	3.02e-6	1.20e-1	4.41e-4
	32	4.04e-13	1.12e-5	1.85e-8	1.15e-8	1.46e-2	2.48e-5
	rate	6.21	3.01	4.03	7.35	2.84	4.05
4	4	4.55e-9	2.71e-4	4.54e-6	2.29e-4	2.12	2.39e-2
	8	2.09e-11	1.55e-5	1.47e-7	6.70e-6	1.38e-1	7.88e-4
	16	1.25e-13	9.38e-7	4.65e-9	9.01e-8	8.72e-3	3.24e-5
	rate	7.58	4.09	4.97	5.65	3.96	4.77

Table 1: Errors and convergence rates for the first and sixth eigenpairs using softFEM and polynomial degrees $p \in \{1,\dots,4\}$.

for the eigenvalues is clearly visible in Figure 1 over the whole spectrum. For the eigenfunctions, there is no difference for $p = 1$ (see Lemma 3.1), whereas the improvement of softFEM over Galerkin FEM for $p \in \{2,3\}$ is salient around the stopping bands (that is, around $j = N^h$ for $p = 2$ and around $j \in \{N^h, 2N^h\}$ for $p = 3$). Incidentally, we notice that for the H^1-seminorm, the errors in the low-frequency region are slightly larger with softFEM than with Galerkin FEM, although the convergence order for softFEM remains optimal. This is expected since in the low-frequency region, best-approximation errors in the finite element space decay optimally, and the softFEM approximation leads to an additional optimally-converging contribution due to the interface jump penalty on the normal gradient. Table 1 reports the errors for the first and sixth eigenpairs using softFEM and polynomial degrees $p \in \{1,\dots,4\}$. We observe that in all the cases, the convergence rates match well the predictions of Theorem 2.3.

To motivate the choice of the softness parameter $\eta = \frac{1}{2(p+1)(p+2)} = \frac{1}{24}$ for $p = 2$,

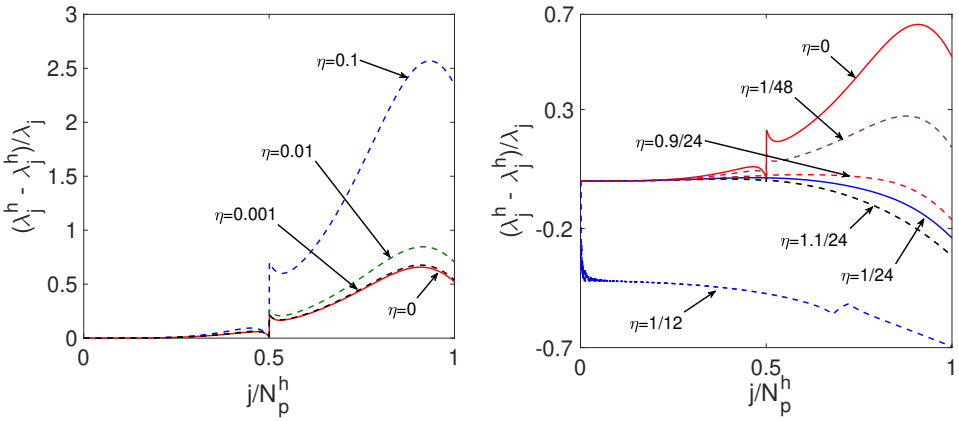

Figure 2: Quadratic softFEM spectra in 1D with $N^h = 1000$ elements using various softness parameters η. Left: $\hat{a} = a + \eta s$; Right: $\hat{a} = a - \eta s$.

we show in Figure 2 the softFEM discrete spectra using various values for the softness parameter η. In this experiment, we increase the mesh resolution to $N^h = 1000$ elements. In the left panel of Figure 2, for the sake of illustration, we actually increase the stiffness, i.e., we set $\hat{a} := a + \eta s$. As expected, increasing η merely worsens the results. Instead, in the right panel of Figure 2, we return to softFEM and consider $\hat{a} := a - \eta s$. We observe that the choice $\eta = \frac{1}{24}$ appears to deliver the best overall result concerning the accuracy of the discrete eigenvalues over the whole spectrum, and that in the high-frequency region, the accuracy of the discrete eigenvalues is sensitive to the value of the softness parameter. For reference, we also display the results for $\eta = \eta_{\max} = \frac{1}{2p(p+1)} = \frac{1}{12}$ which show that the limit value on the softness parameter derived in Theorem 2.2 is indeed sharp.

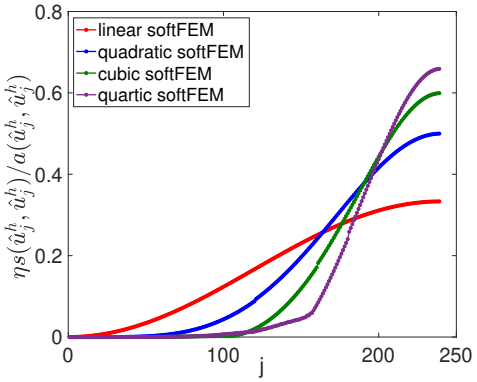

Figure 3: Ratio $\eta s(\hat{u}_j^h, \hat{u}_j^h)/a(\hat{u}_j^h, \hat{u}_j^h)$ for softFEM eigenfunctions. The mesh is composed of 240, 120, 80, 60 uniform elements for $p \in \{1, \ldots, 4\}$, respectively.

In Figure 3 we present the ratio $\eta s(\hat{u}_j^h, \hat{u}_j^h)/a(\hat{u}_j^h, \hat{u}_j^h)$ for softFEM eigenfunctions. The mesh is composed of 240, 120, 80, 60 uniform elements for $p \in \{1, \ldots, 4\}$, respectively.

As predicted by Theorem 2.2, this ratio is always lower than one. We see that the amount of stiffness removed by softFEM is more substantial in the high-frequency region.

p	$\lambda^h_{\min}$	$\lambda^h_{\max}$	$\hat{\lambda}^h_{\max}$	σ	$\hat{\sigma}$	ρ	ϱ
1	9.8698	4.7991e5	3.1995e5	4.8624e4	3.2417e4	1.5000	33.33%
2	9.8696	2.3998e6	1.2000e6	2.4315e5	1.2158e5	1.9999	50.00%
3	9.8696	6.8046e6	2.7255e6	6.8945e5	2.7615e5	2.4967	59.95%
4	9.8696	1.5209e7	5.1587e6	1.5410e6	5.2269e5	2.9482	66.08%
5	9.8696	2.9555e7	9.1006e6	2.9946e6	9.2208e5	3.2476	69.21%

Table 2: Minimal and maximal eigenvalues, condition numbers, stiffness reduction ratios, and percentages when using Galerkin FEM and softFEM for a mesh composed of $N^h = 200$ uniform elements and polynomial degrees $p \in \{1, \dots, 5\}$.

Table 2 shows the minimal and maximal eigenvalues, the condition numbers, the stiffness reduction ratios, and the percentages for Galerkin FEM and softFEM for a mesh composed of $N^h = 200$ uniform elements and polynomial degrees $p \in \{1, \dots, 5\}$. (Recall that $\hat{\lambda}^h_{\min} \approx \lambda^h_{\min}$ so that we only show $\lambda^h_{\min}$ in the table.) We observe that the stiffness reduction ratio increases with the polynomial degree, starting at $\rho = 1.5$ for $p = 1$ up to $\rho = 3.2476$ for $p = 5$. Thus, the benefit of using softFEM in tempering the condition number of the stiffness matrix becomes more pronounced as p is increased. We also notice that the computed value for the stiffness reduction ratio ρ is quite close to the optimal value $1 + \frac{p}{2}$ resulting from Theorem 2.5 (see the lower bound in (2.15)).

3.3 Discrete spectrum for Galerkin FEM

The goal of this section is to briefly outline some basic facts about the spectrum of Galerkin FEM for the 1D Laplace eigenvalue problem. We explore the polynomial degrees $p \in \{1, 2, 3\}$. For $p = 1$, all the degrees of freedom (dofs) in V^h_p are attached to the N^h_p mesh vertices. Letting $\Lambda := \lambda h^2$, solving the GMEVP leads us to look for nonzero vectors in the kernel of the following matrix of order $(N^h - 1) \times (N^h - 1)$:

$$\mathbf{A}_{vv} := \Lambda \begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{bmatrix} - 6 \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}. \tag{3.8}$$

For $p = 2$, there are $N^h_2 = 2N^h - 1$ dofs, and it is interesting to order first the $N^h - 1$ dofs associated with the mesh vertices and then the N^h dofs associated with the mesh elements and whose associated basis functions are bubble functions supported in a single mesh element. Solving the GMEVP problem leads us to look for nonzero vectors in the kernel of the following matrix whose block decomposition reflects the above partition

into vertex and bubble dofs:

$$\begin{bmatrix} \mathbf{A}_{vv} & \mathbf{0} \\ \mathbf{A}_{bv} & \mathbf{A}_{bb} \end{bmatrix}. \tag{3.9}$$

It turns out that there is one vector in the kernel of $\mathbf{A}_{bb}$ whose bubble dofs oscillate from one cell to the next one, and the corresponding eigenvalue is $\lambda_b = 10h^{-2}$. The other vectors are obtained by considering the kernel of the block $\mathbf{A}_{vv}$ which admits the following structure:

$$\begin{aligned} \mathbf{A}_{vv} \coloneqq \Lambda^2 \begin{bmatrix} 6 & -1 & & & \\ -1 & 6 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 6 & -1 \\ & & & -1 & 6 \end{bmatrix} - 16\Lambda \begin{bmatrix} 13 & 1 & & & \\ 1 & & 13 & 1 & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 13 & 1 \\ & & & 1 & 13 \end{bmatrix} \\ + 240 \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}. \end{aligned} \tag{3.10}$$

Finally, for $p = 3$, there are $N_3^h = 3N^h - 1$ dofs, and one orders first the $N^h - 1$ dofs associated with the mesh vertices and then the $2N^h$ dofs associated with the mesh elements and whose associated basis functions are bubble functions supported in a single mesh element (2 per element). Solving the GMEVP problem leads us to look for nonzero vectors in the kernel of a matrix with the same block-structure as in (3.9), but this time the block $\mathbf{A}_{bb}$ is two times larger. The kernel of $\mathbf{A}_{bb}$ is two-dimensional and the corresponding eigenfunctions are thus composed only of bubble functions. Moreover, we have

$$\begin{aligned} \mathbf{A}_{vv} \coloneqq \Lambda^3 \begin{bmatrix} 8 & 1 & & & \\ 1 & 8 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 8 & 1 \\ & & & 1 & 8 \end{bmatrix} - 30\Lambda^2 \begin{bmatrix} 36 & -1 & & & \\ -1 & 36 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 36 & -1 \\ & & & -1 & 36 \end{bmatrix} \\ + 360\Lambda \begin{bmatrix} 64 & 3 & & & \\ 3 & 64 & 3 & & \\ & \ddots & \ddots & \ddots & \\ & & 3 & 64 & 3 \\ & & & 3 & 64 \end{bmatrix} - 25200 \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}. \end{aligned} \tag{3.11}$$

For all $p \in \{1, 2, 3\}$, one can readily verify that the matrix $\mathbf{A}_{vv}$ has a non-trivial kernel if and only if Λ is a root of the following polynomials (the subscript refers to the polynomial

degree):

$$
\begin{aligned}
f_1(\Lambda) &= (2+\zeta_j)\Lambda - 6(1-\zeta_j),\\
f_2(\Lambda) &= 2(3-\zeta_j)\Lambda^2 - 16(13+2\zeta_j)\Lambda + 480(1-\zeta_j),\\
f_3(\Lambda) &= (4+\zeta_j)\Lambda^3 - 30(18-\zeta_j)\Lambda^2 + 360(32+3\zeta_j)\Lambda - 25200(1-\zeta_j),
\end{aligned}
\tag{3.12}
$$

where $\zeta_j := \cos(\pi t_j)$, $t_j := jh$ and $j \in \{1, \ldots, N^h - 1\}$. By replacing ζ_j by the continuous variable $\zeta := \cos(\pi t)$ with $t \in (0,1)$, one obtains one branch of eigenvalues for $p = 1$, two branches of eigenvalues for $p = 2$, and three branches of eigenvalues for $p = 3$. Each branch contains $N^h - 1$ eigenvalues. For $p \in \{2,3\}$, the spectrum is completed by the one or two eigenvalues associated with the eigenfunction(s) composed of bubble functions only.

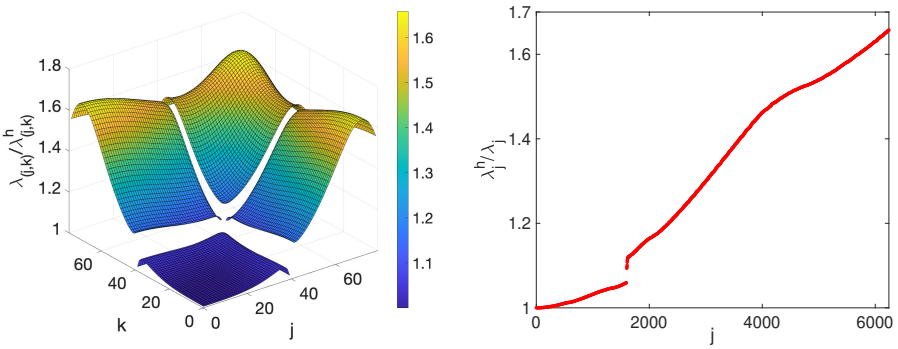

Figure 4: Quadratic FEM approximate spectrum with $N^h = 40 \times 40$ elements for the 2D Laplace eigenvalue problem. Left: eigenvalues sorted in each dimension. Right: eigenvalues sorted in 2D.

In the literature, one refers to these latter eigenvalues as stopping band(s), whereas the branch associated with the lowest eigenvalues is called acoustical branch and the other branches are called optical branches. For instance, [6] reported that quadratic finite elements for the 1D Laplace eigenvalue problem delivered an acoustical branch (low-frequency region) and an optical branch (high-frequency region) separated by one stopping band. We refer the reader to the left plots in Figure 1 for an illustration of these notions. We also observe that the notions of acoustical and optical branches as well as stopping bands depend on the sorting of the eigenvalues and that some overlap between the branches can happen in multiple dimensions; see Figure 4 for an illustration in 2D.

4 SoftFEM on more challenging numerical examples

In this section, we present more challenging numerical tests to illustrate the performances of softFEM. We consider Laplace eigenvalue problems on tensor-product meshes

in Section 4.1, elliptic eigenvalue problems and non-uniform meshes in 1D in Section 4.2, and finally simplicial meshes and L-shaped domains in Section 4.3. The exact eigenpairs of the Laplace eigenvalue problems are known in Section 3 for 1D and Section 4.1 for 2D and 3D, whereas for problems in Sections 4.2 and 4.3 we use a higher-order method with large number of elements to produce reference eigenpairs so as to quantify the approximation errors.

4.1 Laplace eigenvalue problems on tensor-product meshes

We consider the spectral problem (2.1) posed on $\Omega = (0,1)^d$, $d \in \{2,3\}$, with $\kappa = 1$. For $d = 2$, the exact eigenvalues and eigenfunctions are respectively for all $i, j = 1, 2, \ldots$,

$$\lambda_{ij} = (i^2 + j^2)\pi^2, \qquad u_{ij}(x,y) = c_{ij} \sin(i\pi x) \sin(j\pi y),$$

for some normalization constant $c_{ij} > 0$, whereas for $d = 3$, the exact eigenvalues and eigenfunctions are respectively for all $k, l, m = 1, 2, \ldots$,

$$\lambda_{klm} = (k^2 + l^2 + m^2)\pi^2, \quad u_{klm}(x,y,z) = c_{klm} \sin(k\pi x) \sin(l\pi y) \sin(m\pi z),$$

for some normalization constant $c_{klm} > 0$. For the Galerkin FEM and softFEM approximation, we use uniform tensor-product meshes. Theorem 2.2 shows that admissible values for the softness parameter are $\eta \in [0, \eta_{\max})$ with $\eta_{\max} = \frac{1}{2p(p+1)}$. Motivated by the 1D numerical experiments reported Section 3, we take again $\eta = \frac{1}{2(p+1)(p+2)}$.

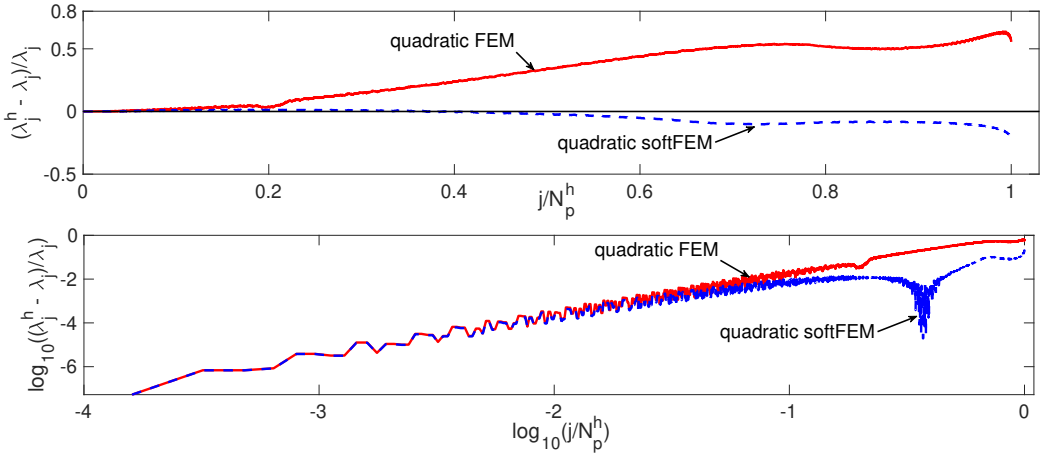

Figure 5: Relative eigenvalue errors for the 2D Laplace eigenvalue problem when using quadratic Galerkin FEM and softFEM with 40^2 elements.

Figures 5 and 6 show the relative eigenvalue errors when using quadratic and cubic Galerkin FEM and softFEM in 2D. For quadratic elements, we use a uniform mesh with 40×40 elements, whereas for cubic elements, we use a uniform mesh with 20×20 elements. Figure 7 shows the eigenvalue errors for the 3D problem with $20 \times 20 \times 20$ elements and

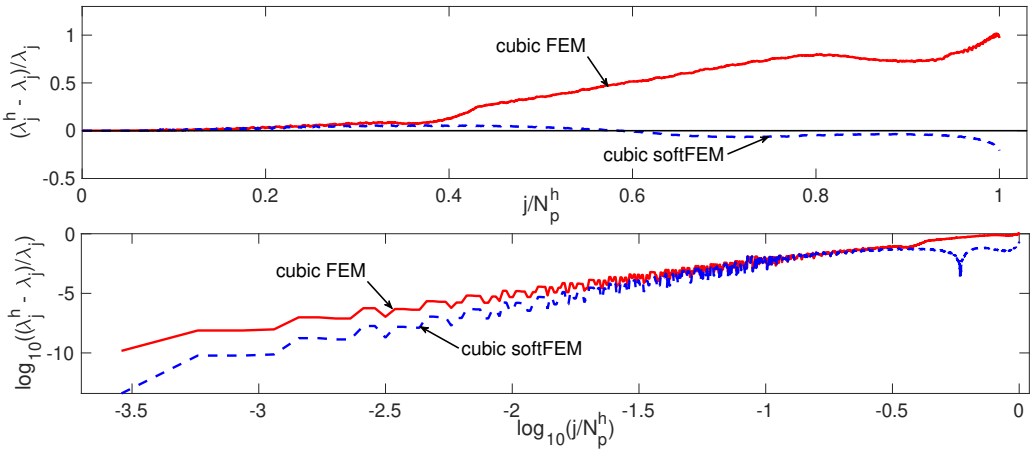

Figure 6: Relative eigenvalue errors for the 2D Laplace eigenvalue problem when using cubic Galerkin FEM and softFEM with 20^2 elements.

$p \in \{2, 3, 4\}$. We observe in these plots that softFEM significantly improves the accuracy in the high-frequency region. Moreover, the plots using the log-log scale indicate that the spectral accuracy is maintained for quadratic elements and even improved for cubic elements in the low-frequency region. The convergence rates for the errors are optimal, and we omit them here for brevity.

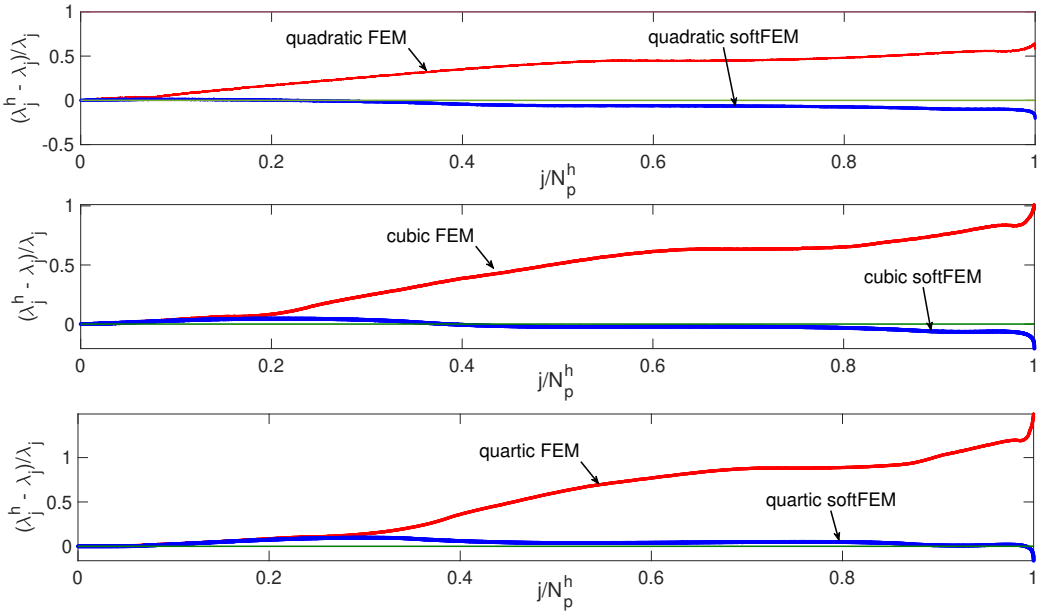

Figure 7: Relative eigenvalue errors for the 3D Laplace eigenvalue problem when using FEM and softFEM with $p = 2, 3, 4$.

Figure 8 shows the ratio $\eta s(\hat{u}_j^h, \hat{u}_j^h)/a(\hat{u}_j^h, \hat{u}_j^h)$ for the softFEM eigenfunctions in

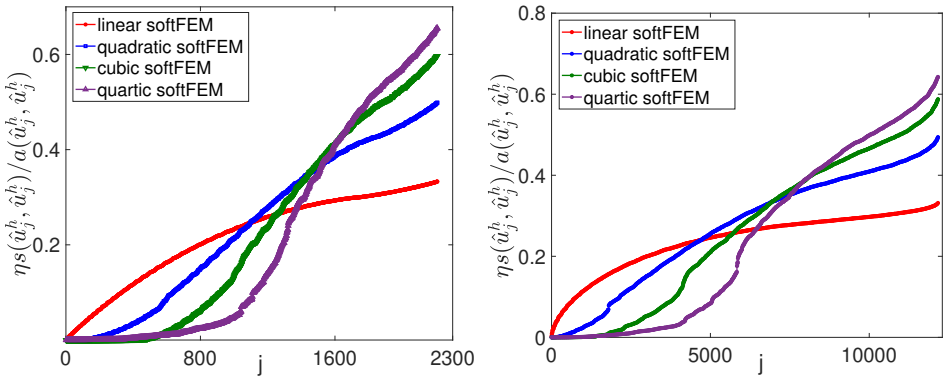

Figure 8: Ratio $\eta s(\hat{u}_j^h, \hat{u}_j^h)/a(\hat{u}_j^h, \hat{u}_j^h)$ for the softFEM eigenfunctions for the Laplace eigenvalue problem in 2D (left) and 3D (right).

both the 2D and 3D settings. In 2D, there are 48×48, 24×24, 16×16, and 12×12 uniform elements for $p \in \{1, \ldots, 4\}$, respectively, whereas in 3D, there are $24 \times 24 \times 24$, $12 \times 12 \times 12$, $8 \times 8 \times 8$, and $6 \times 6 \times 6$ uniform elements for $p \in \{1, \ldots, 4\}$, respectively. These results show essentially how much stiffness is removed from the eigenfunctions by means of softFEM. The fact that the ratio $\eta s(\hat{u}_j^h, \hat{u}_j^h)/a(\hat{u}_j^h, \hat{u}_j^h)$ is more pronounced in the high-frequency region corroborates the reduction of the spectral errors in this region. Finally, we mention that the stiffness reduction ratios and percentages are quite close to those reported in 1D, that is, $\rho \approx 1 + \frac{p}{2}$ and $\varrho \approx 100\frac{p}{p+2}\%$ for $p \in \{1, \ldots, 4\}$ in both 2D and 3D.

4.2 Elliptic eigenvalue problems and non-uniform meshes in 1D

We now consider the 1D elliptic eigenvalue problem (2.1) with $\kappa(x) := e^{x \sin(2\pi x)}$, so that $\kappa_{\max} \approx 1.34$ and $\kappa_{\min} \approx 0.46$. The exact eigenpairs are approximated using Galerkin FEM with C^6 septic B-spline basis functions and a mesh composed of $N^h = 1000$ elements. Figure 9 compares the relative eigenvalue errors for Galerkin FEM and softFEM on a uniform mesh composed of $N^h = 200$ elements and polynomial degrees $p \in \{2, 3, 4, 5\}$. We observe that softFEM reduces the spectrum errors, especially in the high-frequency region. The convergence rates for the errors are optimal, and we omit them here for brevity.

Table 3 shows the smallest and largest eigenvalues, the condition numbers, the stiffness reduction ratios, and the percentages for Galerkin FEM and softFEM. In all cases, we observe that softFEM leads to smaller largest eigenvalues and hence to smaller condition numbers. The stiffness reduction ratio is about $\rho = \sigma/\hat{\sigma} \approx 1 + \frac{p}{2}$ while the percentage is about $\varrho \approx 100\frac{p}{p+2}\%$; this is consistent with the 1D results reported in Section 3.2.

Figure 10 compares the relative eigenvalue errors for the 1D Laplace eigenvalue problem when using Galerkin FEM and softFEM with $p \in \{2, 3, 4, 5\}$ on a non-uniform mesh composed of $N^h = 10$ elements. The mesh nodes have been randomly set to

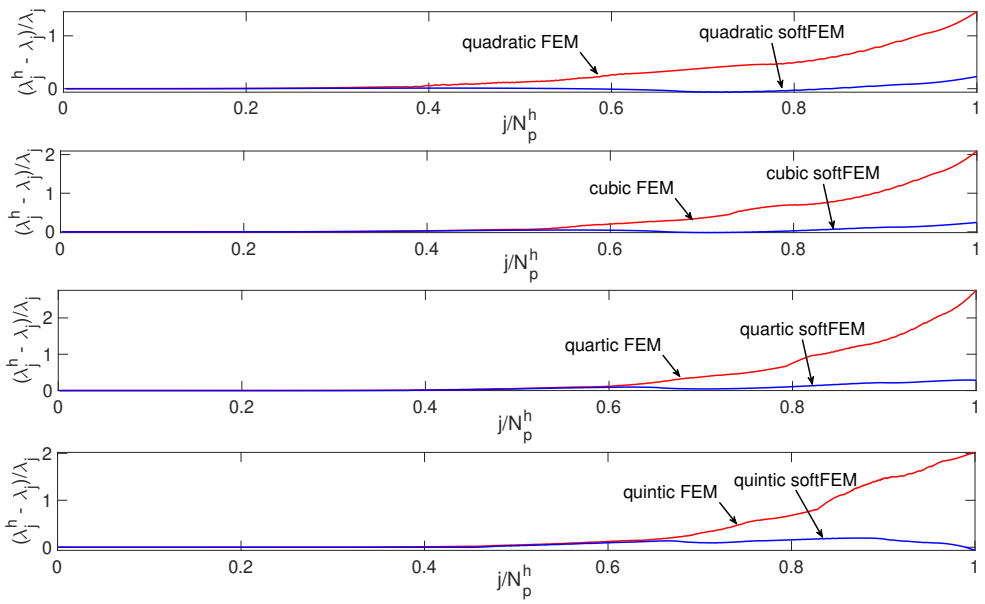

Figure 9: Relative eigenvalue errors for the elliptic eigenvalue problem (2.1) in 1D with $\kappa(x) := e^{x\sin(2\pi x)}$ when using Galerkin FEM and softFEM with $p \in \{2,3,4,5\}$. The mesh has $N^h = 200$ uniform elements.

$\{0, 0.1, 0.18, 0.29, 0.41, 0.5, 0.59, 0.66, 0.81, 0.92, 1\}$. Reference eigenvalues to evaluate the errors are computed as above. We observe that the improvement offered by softFEM over Galerkin FEM is similar to the one observed on uniform meshes. Table 4 reports the smallest and largest eigenvalues, the condition numbers, the stiffness reduction ratios, and the percentages. We observe that the stiffness reduction ratios are slightly larger than when using uniform meshes (compare with Table 3).

4.3 Simplicial meshes and L-shaped domain

In this section we consider the 2D Laplace eigenvalue problem posed on the unit square domain or on the L-shaped domain, and we use simplicial meshes (triangulations) as depicted in Figure 11. Theorem 2.2 shows that admissible values for the softness parameter on simplicial meshes are $\eta \in [0, \eta_{\max})$ with $\eta_{\max} = \frac{1}{2p(p+d-1)} = \frac{1}{2p(p+1)}$ if $d = 2$. Motivated by the 1D numerical experiments reported in the previous section, we take again $\eta = \frac{1}{2(p+1)(p+2)}$.

Figures 12 and 13 compare the relative eigenvalue errors for the 2D Laplace eigenvalue problem on the unit square domain and the L-shaped domain, respectively, when using Galerkin FEM and softFEM with $p \in \{1,2,3\}$ and an unstructured mesh (triangulation). As observed in the previous numerical experiments, softFEM leads to smaller spectral errors than Galerkin FEM especially in the high-frequency region.

Tables 5 and 6 report the smallest and largest eigenvalues, the condition numbers, the stiffness reduction ratios and the percentages for the 2D Laplace eigenvalue problem

p	$\lambda^h_{\min}$	$\lambda^h_{\max}$	$\hat{\lambda}^h_{\max}$	σ	$\hat{\sigma}$	ρ	ϱ
1	8.2832	6.3326e5	4.2263e5	7.6451e4	5.1023e4	1.4984	33.26%
2	8.2829	3.1795e6	1.5936e6	3.8386e5	1.9240e5	1.9951	49.88%
3	8.2829	9.0280e6	3.6298e6	1.0900e6	4.3823e5	2.4872	59.79%
4	8.2829	2.0194e7	6.8865e6	2.4380e6	8.3141e5	2.9323	65.90%
5	8.2829	3.9263e7	1.2129e7	4.7402e6	1.4643e6	3.2371	69.11%

Table 3: Minimal and maximal eigenvalues, condition numbers, stiffness reduction ratios, and percentages for the 1D elliptic eigenvalue problem with $\kappa(x) := e^{x\sin(2\pi x)}$ when using Galerkin FEM and softFEM on a uniform mesh composed of $N^h = 200$ elements.

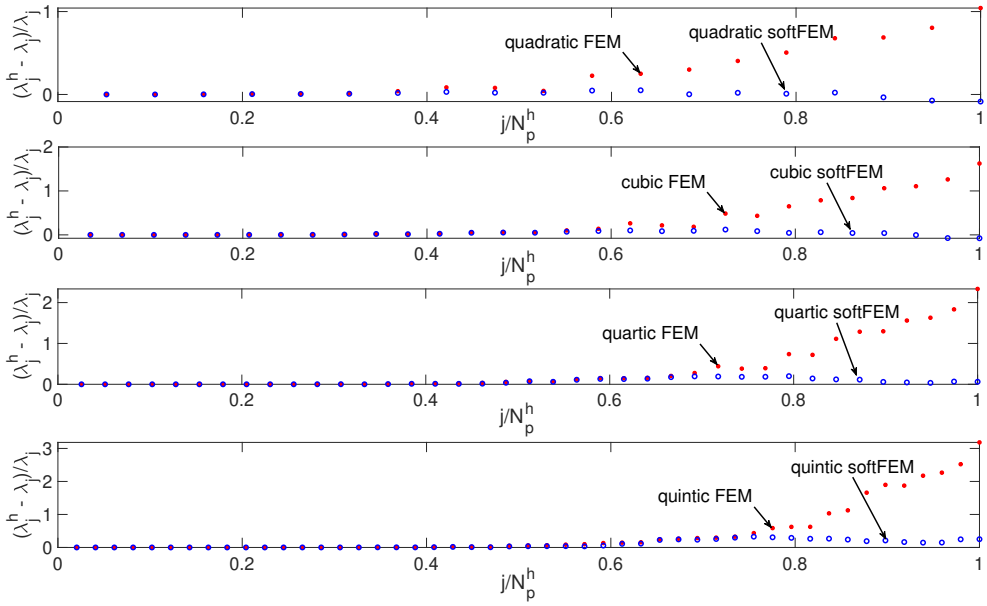

Figure 10: Relative eigenvalue errors for the 1D Laplace eigenvalue problem when using Galerkin FEM and softFEM with $p \in \{2,3,4,5\}$ on a non-uniform mesh composed of $N^h = 10$ elements. The mesh nodes have been randomly set to $\{0, 0.1, 0.18, 0.29, 0.41, 0.5, 0.59, 0.66, 0.81, 0.92, 1\}$.

on the unit square domain and the L-shaped domain respectively when using Galerkin FEM and softFEM with $p \in \{1,2,3\}$ and an unstructured mesh. Once again we observe that softFEM is capable to reduce significantly the stiffness of the resulting matrix on unstructured meshes as well.

p	$\lambda^h_{\min}$	$\lambda^h_{\max}$	$\hat{\lambda}^h_{\max}$	σ	$\hat{\sigma}$	ρ	ϱ
1	9.9653	1.2631e3	8.0985e2	1.2675e2	8.1267e1	1.5597	35.88%
2	9.8698	7.2767e3	3.2585e3	7.3727e2	3.3014e2	2.2332	55.22%
3	9.8696	2.1782e4	7.6596e3	2.2070e3	7.7608e2	2.8438	64.84%
4	9.8696	5.0056e4	1.5948e4	5.0717e3	1.6159e3	3.1387	68.14%
5	9.8696	9.9119e4	2.9618e4	1.0043e4	3.0009e3	3.3466	70.12%

Table 4: Minimal and maximal eigenvalues, condition numbers, stiffness reduction ratios, and percentages for the 1D Laplace eigenvalue problem when using Galerkin FEM and softFEM on a non-uniform mesh composed of $N^h = 10$ elements. The mesh nodes have been randomly set to $\{0, 0.1, 0.18, 0.29, 0.41, 0.5, 0.59, 0.66, 0.81, 0.92, 1\}$.

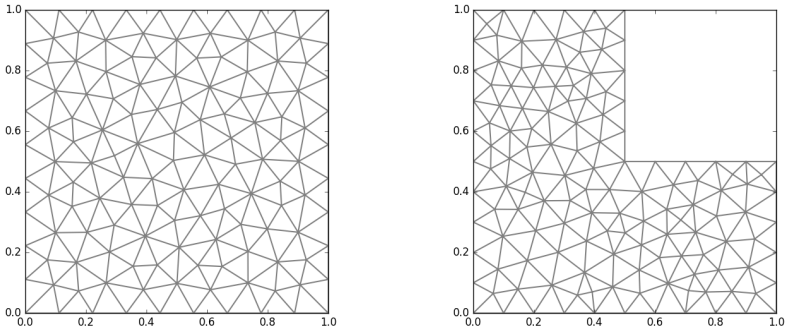

Figure 11: Unstructured meshes for the unit square domain (left) and the L-shaped domain (right).

p	$\lambda^h_{\min}$	$\lambda^h_{\max}$	$\hat{\lambda}^h_{\max}$	σ	$\hat{\sigma}$	ρ	ϱ
1	2.0020e1	2.9992e3	9.8013e2	1.4981e2	4.8957e1	3.0600	67.32%
2	1.9740e1	1.5224e4	4.1819e3	7.7122e2	2.1185e2	3.6404	72.53%
3	1.9739e1	4.0719e4	1.2356e4	2.0628e3	6.2598e2	3.2954	69.65%

Table 5: Minimal and maximal eigenvalues, condition numbers, stiffness reduction ratios, and percentages for the 2D Laplace eigenvalue problem on the unit square domain when using Galerkin FEM and softFEM with $p \in \{1, 2, 3\}$ and an unstructured mesh.

5 Proof of Theorem 2.2

In this section we prove Theorem 2.2 which establishes the coercivity of the bilinear form $\hat{a}(\cdot, \cdot)$ under the condition that the softness parameter $\eta \in [0, \eta_{\max})$ for some real number $\eta_{\max}$ depending on the polynomial degree p and the type of mesh. To this purpose we first establish some useful discrete trace inequalities.

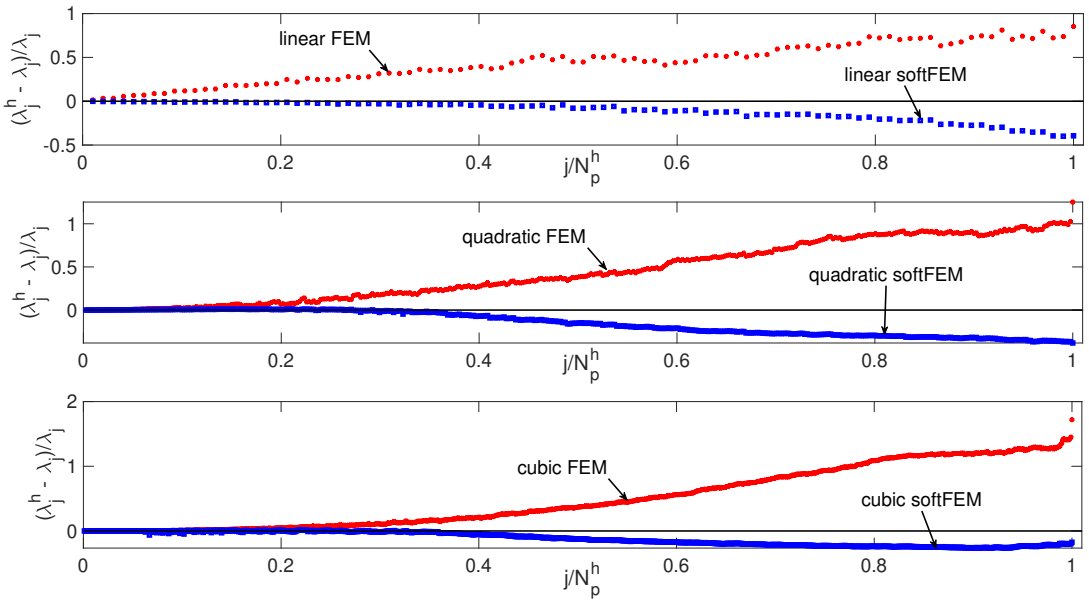

Figure 12: Relative eigenvalue errors for the 2D Laplace eigenvalue problem on the unit square domain when using Galerkin FEM and softFEM with $p \in \{1, 2, 3\}$ and an unstructured mesh.

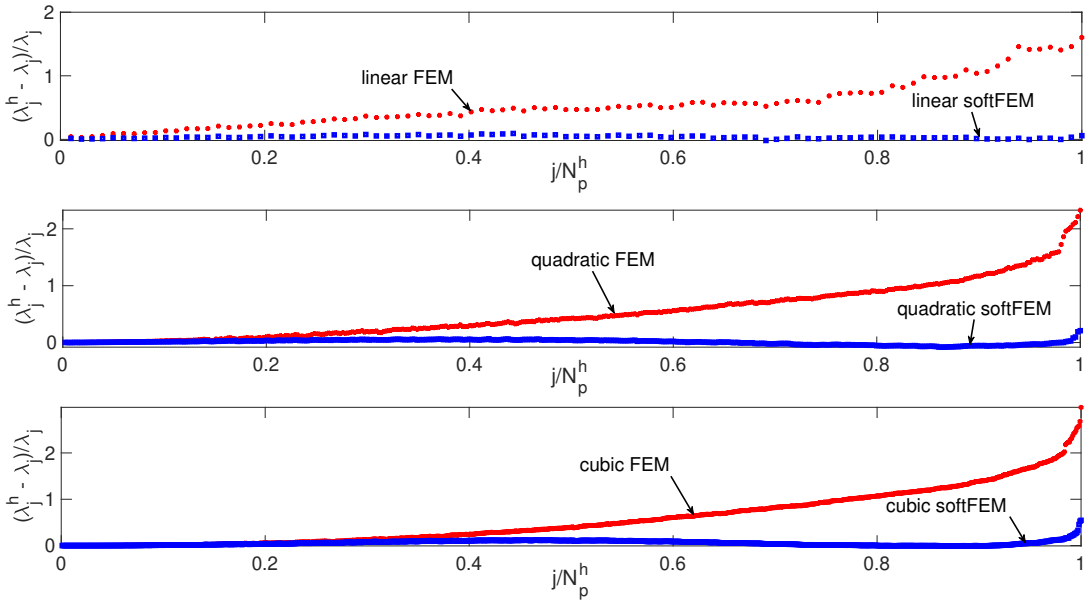

Figure 13: Relative eigenvalue errors for the 2D Laplace eigenvalue problem on the L-shaped domain when using Galerkin FEM and softFEM with $p \in \{1, 2, 3\}$ and an unstructured mesh.

p	$\lambda_{\min}^h$	$\lambda_{\max}^h$	$\hat{\lambda}_{\max}^h$	σ	$\hat{\sigma}$	ρ	ϱ
1	4.0162e1	4.7287e3	1.9228e3	1.1774e2	4.7875e1	2.4593	59.34%
2	3.8707e1	2.5394e4	9.2240e3	6.5605e2	2.3830e2	2.7530	63.68%
3	3.8619e1	7.1172e4	2.7611e4	1.8429e3	7.1496e2	2.5777	61.21%

Table 6: Minimal and maximal eigenvalues, condition numbers, stiffness reduction ratios, and percentages for the 2D Laplace eigenvalue problem on the L-shaped domain when using Galerkin FEM and softFEM with $p \in \{1,2,3\}$ and an unstructured mesh.

5.1 Discrete trace inequalities

For a natural number $m \in \mathbb{N}$, we define the sets $I_m := \{0,\dots,m\}$, $\partial I_m := \{0,m\}$, and $I_m^0 := I_m \setminus \partial I_m$. Let $p \in \mathbb{N}$ be the polynomial degree. We are going to consider the Gauss–Lobatto rule with $(p+2)$ points (see, for example, [30]), which is exact for polynomials of degree at most $(2p+1)$. The weights are denoted $\{\varpi_j\}_{j\in I_{p+1}}$ and the nodes in $[-1,1]$ are denoted $\{\xi_j\}_{j\in I_{p+1}}$. Recall that

$$\varpi_0 = \varpi_{p+1} = \frac{2}{(p+1)(p+2)}, \qquad \varpi_j = \frac{2}{(p+1)(p+2)(L_{p+1}(\xi_j))^2}, \quad \forall j \in I_{p+1}^0, \tag{5.1}$$

where L_{p+1} is the Legendre polynomial of degree $(p+1)$. For a univariate function v that is k-times differentiable, we denote its k-th derivative as $v^{(k)}$.

Lemma 5.1 (Discrete trace inequality, 1D)**.** *Let $\tau := [a,b]$ with $b > a$ and $\partial\tau = \{a,b\}$. Set $h_\tau := b-a$. For all $p \in \mathbb{N}$ and all $k \in I_p$, the following holds:*

$$\|v^{(k)}\|_{\partial\tau} \le C_1(k,p) h_\tau^{-1/2} \, \|v^{(k)}\|_\tau, \quad \forall v \in \mathbb{P}_p(\tau), \tag{5.2}$$

where $C_1(k,p) := \sqrt{(p-k+1)(p-k+2)}$. Moreover, the constant $C_1(k,p)$ is sharp. In particular, for $p \ge 1$ and $k = 1$, we have

$$\|v'\|_{\partial\tau} \le \sqrt{p(p+1)} h_\tau^{-1/2} \, \|v'\|_\tau, \quad \forall v \in \mathbb{P}_p(\tau). \tag{5.3}$$

Proof. It is clear that it suffices to prove (5.2) for $k = 0$. Let $v \in \mathbb{P}_p(\tau)$. We observe that $\|v\|_{\partial\tau}^2 = v(a)^2 + v(b)^2$. Moreover, since v^2 is a polynomial of degree at most $2p$, it is integrated exactly by the Gauss–Lobatto quadrature with $(p+2)$ points. Considering the linear mapping from $\xi \in [-1,1]$ to $[a,b]$ with $x(\xi) := \frac{b-a}{2}\xi + \frac{a+b}{2}$ and setting $g_j := x(\xi_j)$ for all $j \in I_{p+1}$, we have

$$\begin{aligned}
\|v\|_\tau^2 &= \int_a^b v^2(x)\, dx = \int_{-1}^1 v^2(x(\xi)) \frac{dx}{d\xi}\, d\xi \\
&= \frac{b-a}{2}\left(\frac{2v^2(a) + 2v^2(b)}{(p+1)(p+2)} + \sum_{j\in I_{p+1}^0} \varpi_j v^2(g_j) \right) \\
&= \frac{b-a}{(p+1)(p+2)} \|v\|_{\partial\tau}^2 + \frac{b-a}{2} \sum_{j\in I_{p+1}^0} \varpi_j v^2(g_j) \ge C_1(0,p)^{-2} h_\tau \|v\|_{\partial\tau}^2,
\end{aligned}$$

where we used that $g_0 = a$ and $g_{p+1} = b$, the definition of $C_1(0,p)$ and h_τ, and the fact that the weights ϖ_j are non-negative for all $j \in I^0_{p+1}$. This proves (5.1) for $k = 0$. Finally, that the inequality is sharp follows from the fact that it is possible to find a nonzero polynomial in $\mathbb{P}_p(\tau)$ that vanishes at all the points g_j for all $j \in I^0_{p+1}$. □

Let us now turn to the multi-dimensional case. We consider first the tensor-product case. For simplicity we focus on bounding the normal derivative on the boundary of a cuboid cell. For a different result bounding any partial derivative on the boundary, we refer the reader to Remark 5.4.

Lemma 5.2 (Discrete trace inequality, cuboid)**.** *Let $\tau := [a_1, b_1] \times \ldots \times [a_d, b_d] \subset \mathbb{R}^d$, with $b_j > a_j$ for all $j \in \{1, \ldots, d\}$, be a cuboid with boundary $\partial\tau$ and outward normal $\boldsymbol{n}_\tau$. Recall that $h^0_\tau := \min_{i\in\{1,\ldots,d\}}(b_i - a_i)$ is the length of the smallest edge of τ. Let $p \geq 1$. The following holds:*

$$\|\nabla v \cdot \boldsymbol{n}_\tau\|_{\partial\tau} \leq \sqrt{p(p+1)}(h^0_\tau)^{-1/2}\|\nabla v\|_\tau, \quad \forall v \in \mathbb{Q}_p(\tau), \tag{5.4}$$

Moreover, the constant is sharp. Notice that (5.4) *coincides with* (5.3) *for $d = 1$.*

Proof. We present the proof in the 2D case ($d = 2$); the general case is treated similarly. Let $v \in \mathbb{Q}_p(\tau)$. One can write $v(x,y) = \sum_{j_x,j_y\in I_p} \alpha_{j_xj_y}\psi_{j_x}(x)\psi_{j_y}(y)$, where $\{\psi_j\}_{j\in I_p}$ are basis functions of the univariate polynomial space of degree at most p. Moreover, we have $\partial\tau = \mathcal{F}_x \cup \mathcal{F}_y$. $\mathcal{F}_x$ contains two faces (located at $x = a_1, b_1$) and so does $\mathcal{F}_y$ (located at $y = a_2, b_2$). We consider the linear mappings $x : [-1,1] \to [a_1, b_1]$ and $y : [-1,1] \to [a_2, b_2]$. Let us first consider the two faces in $\mathcal{F}_x$. Since we are integrating the partial derivative of v with respect to x, we consider a Gauss–Lobatto quadrature in τ obtained as the tensor-product of a Gauss–Lobatto quadrature with $(p+1)$ points in the x variable and a Gauss–Lobatto quadrature with $(p+2)$ points in the y variable. We use a superscript for the weights and nodes to indicate the number of points in the quadrature, and we set $g^x_j := x(\xi^{p+1}_j)$ for all $j \in I_p$ and $g^y_j := y(\xi^{p+2}_j)$ for all $j \in I_{p+1}$. Using the same arguments as in the proof of Lemma 5.1, we obtain

$$\begin{aligned}
\|\nabla v \cdot \boldsymbol{n}_\tau\|^2_{\mathcal{F}_x} &= \int_{a_2}^{b_2} (\partial_x v|_{x=a_1})^2 \, dy + \int_{a_2}^{b_2} (\partial_x v|_{x=b_1})^2 \, dy \\
&= \frac{b_2 - a_2}{2} \sum_{l_x\in\partial I_p} \sum_{l_y\in I_{p+1}} \varpi^{p+2}_{l_y} \left(\sum_{j_x,j_y\in I_p} \alpha_{j_xj_y}\psi'_{j_x}(g^x_{l_x})\psi_{j_y}(g^y_{l_y}) \right)^2 \\
&= \frac{b_2 - a_2}{2}\frac{p(p+1)}{2} \sum_{l_x\in\partial I_p} \sum_{l_y\in I_{p+1}} \varpi^{p+1}_{l_x}\varpi^{p+2}_{l_y} \left(\sum_{j_x,j_y\in I_p} \alpha_{j_xj_y}\psi'_{j_x}(g^x_{l_x})\psi_{j_y}(g^y_{l_y}) \right)^2 \\
&\leq \frac{b_2 - a_2}{2}\frac{p(p+1)}{2} \sum_{l_x\in I_p, l_y\in I_{p+1}} \varpi^{p+1}_{l_x}\varpi^{p+2}_{l_y} \left(\sum_{j_x,j_y\in I_p} \alpha_{j_xj_y}\psi'_{j_x}(g^x_{l_x})\psi_{j_y}(g^y_{l_y}) \right)^2 \\
&= \frac{p(p+1)}{b_1 - a_1}\|\partial_x v\|^2_\tau.
\end{aligned}$$

Similarly we have

$$\|\nabla v \cdot \boldsymbol{n}_\tau\|^2_{\mathcal{F}_y} \le \frac{p(p+1)}{b_2-a_2}\|\partial_y v\|^2_\tau.$$

Summing the above two inequalities and recalling the definition of h^0_τ gives

$$\|\nabla v \cdot \boldsymbol{n}_\tau\|^2_{\partial\tau} \le \frac{p(p+1)}{b_1-a_1}\|\partial_x v\|^2_\tau + \frac{p(p+1)}{b_2-a_2}\|\partial_y v\|^2_\tau \le \frac{p(p+1)}{h^0_\tau}\|\nabla v\|^2_\tau. \tag{5.5}$$

Taking square roots completes the proof for $d = 2$. Finally, the constant is sharp since the upper bound in (5.4) can be attained by univariate functions. □

Finally we consider the case of a simplex.

Lemma 5.3 (Discrete trace inequality, simplex)**.** *Let τ be a simplex in $\mathbb{R}^d$, $d \ge 2$, with boundary $\partial\tau$ and outward normal $\boldsymbol{n}_\tau$. Recall that $h^0_\tau := \frac{d|\tau|}{|\partial\tau|}$. Let $p \ge 1$. The following holds:*

$$\|\nabla v \cdot \boldsymbol{n}_\tau\|_{\partial\tau} \le \sqrt{p(p+d-1)}(h^0_\tau)^{-1/2}\|\nabla v\|_\tau, \quad \forall v \in \mathbb{P}_p(\tau). \tag{5.6}$$

Proof. Let $v \in \mathbb{P}_p(\tau)$, let F be a face of τ and set $\boldsymbol{n}_F := \boldsymbol{n}_{\tau|F}$. Then $w := \nabla v \cdot \boldsymbol{n}_F \in \mathbb{P}_{p-1}(\tau)$. Applying the discrete trace inequality from [34] yields

$$\|w\|^2_{L^2(F)} \le p(p+d-1)\frac{|F|}{d|\tau|}\|w\|^2_{L^2(\tau)}.$$

Since $|\nabla v \cdot \boldsymbol{n}_F| \le \|\nabla v\|_{\ell^2}$ (the Euclidean norm of ∇v), we infer that

$$\|\nabla v \cdot \boldsymbol{n}_F\|^2_{L^2(F)} \le p(p+d-1)\frac{|F|}{d|\tau|}\|\nabla v\|^2_{L^2(\tau)}.$$

Summing over the faces of τ, taking square roots, and recalling the definition of h^0_τ conclude the proof. □

Remark 5.4 (Lemma 5.1)**.** Using the Gauss–Lobatto nodes and their tensor-products to prove discrete trace inequalities is a known technique. The result of Lemma 5.1 however slightly differs from previous results from the literature and provides a sharper constant. For instance, for $p \ge 1$, $d = 1$ and $k = 0$, [34, Thm. 2] leads to the constant $\sqrt{2(p+1)^2}$ and [8, Lemma 3.1] to the constant $\sqrt{(p+1)(2p+1)}$, which are both less sharp than $C_1(0,p)$ in (5.2). Notice also that (5.6) with $d = 1$ leads to $\|v'\|_{\partial\tau} \le \sqrt{2p^2}h_\tau^{-1/2}\|v'\|_\tau$ which is again less sharp that (5.3) for $p \ge 2$. Finally, we have the following multidimensional extension of Lemma 5.1 in a cuboid; the proof is omitted for brevity and follows arguments similar to those above. Let $\tau := [a_1,b_1]\times\ldots\times[a_d,b_d] \subset \mathbb{R}^d$, with $b_j > a_j$ for all $j \in \{1,\ldots,d\}$, be a cuboid with boundary $\partial\tau$. Let $p \ge 1$. For any multi-index $(\boldsymbol{k}) = (k_1,\ldots,k_d)$ with $k_j \in I_p$ for all $j \in \{1,\ldots,d\}$, denoting the $\boldsymbol{k}$-th partial derivative of v as $v^{(\boldsymbol{k})}$, the following holds:

$$\|v^{(\boldsymbol{k})}\|_{\partial\tau} \le C_d(\boldsymbol{k},p,\tau)\ \|v^{(\boldsymbol{k})}\|_\tau, \quad \forall v \in \mathbb{Q}_p(\tau), \tag{5.7}$$

with $C_d(\boldsymbol{k},p,\tau) := \sqrt{\sum_{j\in\{1,\ldots,d\}}\frac{(p-k_j+1)(p-k_j+2)}{b_j-a_j}}$. Moreover, the constant $C_d(\boldsymbol{k},p,\tau)$ is sharp.

5.2 Coercivity proof

We can now give the proof of Theorem 2.2.

Proof of Theorem 2.2. (i) Tensor-product meshes. For all $F \in \mathcal{F}_h^i$, let $\mathcal{T}_F$ be the set collecting the two mesh elements sharing F. For all $w^h \in V_p^h$, we have

$$s(w^h, w^h) = \sum_{F\in\mathcal{F}_h^i} \kappa_F h_F \|[\![\nabla w^h \cdot \boldsymbol{n}]\!]\|_F^2 \le 2 \sum_{F\in\mathcal{F}_h^i} \sum_{\tau\in\mathcal{T}_F} \kappa_F h_F \|\nabla w^h|_\tau \cdot \boldsymbol{n}_\tau\|_F^2.$$

Since $h_F = \min_{\tau\in\mathcal{T}_F} h_\tau^0$ and $\kappa_F = \min_{\tau\in\mathcal{T}_F} \kappa_\tau$ (see (2.8)) and exchanging the order of the two summations, we infer that

$$s(w^h, w^h) \le 2 \sum_{\tau\in\mathcal{T}_h} \kappa_\tau h_\tau^0 \|\nabla w^h \cdot \boldsymbol{n}_\tau\|_{\partial\tau}^2.$$

Applying Lemma 5.2 yields

$$s(w^h, w^h) \le 2p(p+1) \sum_{\tau\in\mathcal{T}_h} \kappa_\tau \|\nabla w^h\|_\tau^2 \le 2p(p+1) a(w^h, w^h).$$

Recalling that $\hat{a}(\cdot,\cdot) = a(\cdot,\cdot) - \eta s(\cdot,\cdot)$ with $\eta > 0$, we conclude that

$$\hat{a}(w^h, w^h) \ge (1 - 2p(p+1)\eta) a(w^h, w^h). \tag{5.8}$$

(ii) Simplicial meshes. The proof is similar but we now invoke Lemma 5.3 instead of Lemma 5.2. □

6 Concluding remarks

In this work, we have shown by mathematical analysis and numerical experiments the benefits of tempering the stiffness of the Galerkin FEM approximation of a model elliptic spectral problem by subtracting to the stiffness bilinear form a least-squares penalty on the gradient jumps across the mesh interfaces. This novel approximation technique has been named softFEM since it reduces the stiffness of the problem. SoftFEM is formulated in terms of one softness parameter for which we provided an admissible range of values to maintain coercivity on both tensor-product and simplicial meshes, as well as a practical choice for its value that leads to superconvergence for linear softFEM in 1D and to attractive numerical performances in more general situations. The main feature of softFEM is that it preserves the optimal accuracy of the eigenvalues in the low-frequency region, while at the same time improving significantly the accuracy in the high-frequency region. The main explanation for this improvement is, as illustrated numerically in our experiments, that in the high-frequency region the standard Galerkin FEM approximation tends to store a substantial amount of energy for the eigenfunctions in the form of gradient jumps across the mesh interfaces. Another very important

advantage of softFEM that we illustrated in several settings is its ability to offer a sizable reduction of the conditioning of the stiffness matrix. The optimal value of the asymptotic stiffness reduction ratio increases linearly with the polynomial degree and fairly close values to those predicted theoretically are recovered in our various numerical experiments.

As for future work, a first possible direction is the generalization to other operators and to other FEM-based methods, such as FEM with C^1 cubic elements and discontinuous Galerkin methods for instance. Moreover, the stiffness reduction by softFEM lends itself naturally to tempering the CFL condition in explicit time-marching schemes applied to time-dependent PDEs and to the approximation of the Helmholtz equation.

References

[1] P. F. Antonietti, A. Buffa, and I. Perugia, *Discontinuous Galerkin approximation of the Laplace eigenproblem*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3483–3503.

[2] I. Babuška and J. Osborn, *Eigenvalue problems*, in Handbook of Numerical Analysis, Vol. II, Handb. Numer. Anal., II, North-Holland, Amsterdam, 1991, pp. 641–787.

[3] D. Boffi, *Finite element approximation of eigenvalue problems*, Acta Numer., 19 (2010), pp. 1–120.

[4] J. H. Bramble and J. E. Osborn, *Rate of convergence estimates for nonselfadjoint eigenvalue approximations*, Math. Comp., 27 (1973), pp. 525–549.

[5] H. Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Universitext, Springer, New York, 2011.

[6] L. Brillouin, *Wave propagation in periodic structures*, Dover Publications, Inc., (1953).

[7] E. Burman, *A unified analysis for conforming and nonconforming stabilized finite element methods using interior penalty*, SIAM J. Numer. Anal., 43 (2005), pp. 2012–2033.

[8] E. Burman and A. Ern, *Continuous interior penalty hp-finite element methods for advection and advection-diffusion equations*, Math. Comp., 76 (2007), pp. 1119–1140.

[9] E. Burman and P. Hansbo, *Edge stabilization for the generalized Stokes problem: a continuous interior penalty method*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 2393–2410.

[10] V. Calo, M. Cicuttin, Q. Deng, and A. Ern, *Spectral approximation of elliptic operators by the hybrid high-order method*, Math. Comp., 88 (2019), pp. 1559–1586.

[11] V. Calo, Q. Deng, and V. Puzyrev, *Dispersion optimized quadratures for isogeometric analysis*, J. Comput. Appl. Math., 355 (2019), pp. 283–300.

[12] C. Canuto, *Eigenvalue approximations by mixed methods*, RAIRO Anal. Numér., 12 (1978), pp. 27–50, iii.

[13] C. Carstensen, A. Ern, and S. Puttkammer, *Guaranteed lower bounds on eigenvalues of elliptic operators with a hybrid high-order method*, hal.archives-ouvertes, (2020). available at `https://hal.archives-ouvertes.fr/hal-02863599`.

[14] B. Cockburn, J. Gopalakrishnan, F. Li, N.-C. Nguyen, and J. Peraire, *Hybridization and postprocessing techniques for mixed eigenfunctions*, SIAM J. Numer. Anal., 48 (2010), pp. 857–881.

[15] J. A. Cottrell, A. Reali, Y. Bazilevs, and T. J. R. Hughes, *Isogeometric analysis of structural vibrations*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 5257–5296.

[16] Q. Deng, *Analytical solutions to some generalized and polynomial eigenvalue problems*, arXiv preprint arXiv:2007.08130, (2020).

[17] Q. Deng and V. Calo, *Dispersion-minimized mass for isogeometric analysis*, Comput. Methods Appl. Mech. Engrg., 341 (2018), pp. 71–92.

[18] ——, *A boundary penalization technique to remove outliers from isogeometric analysis on tensor-product meshes*, arXiv preprint arXiv:2010.08159, (2020).

[19] J. Descloux, N. Nassif, and J. Rappaz, *On spectral approximation. I. The problem of convergence*, RAIRO Anal. Numér., 12 (1978), pp. 97–112, iii.

[20] ——, *On spectral approximation. II. Error estimates for the Galerkin method*, RAIRO Anal. Numér., 12 (1978), pp. 113–119, iii.

[21] A. Ern and J.-L. Guermond, *Finite Elements II: Galerkin approximation, elliptic and mixed PDEs*, Springer-Verlag, New York, 2020. In press.

[22] F. Gardini and G. Vacca, *Virtual element method for second-order elliptic eigenvalue problems*, IMA J. Numer. Anal., 38 (2018), pp. 2026–2054.

[23] S. Giani, *hp-adaptive composite discontinuous Galerkin methods for elliptic eigenvalue problems on complicated domains*, Appl. Math. Comput., 267 (2015), pp. 604–617.

[24] J. Gopalakrishnan, F. Li, N.-C. Nguyen, and J. Peraire, *Spectral approximations by the HDG method*, Math. Comp., 84 (2015), pp. 1037–1059.

[25] T. J. R. Hughes, A. Reali, and G. Sangalli, *Duality and unified analysis of discrete approximations in structural dynamics and wave propagation: comparison of p-method finite elements with k-method NURBS*, Comput. Methods Appl. Mech. Engrg., 197 (2008), pp. 4104–4124.

[26] B. Mercier, J. Osborn, J. Rappaz, and P.-A. Raviart, *Eigenvalue approximation by mixed and hybrid methods*, Math. Comp., 36 (1981), pp. 427–453.

[27] B. Mercier and J. Rappaz, *Eigenvalue approximation via non-conforming and hybrid finite element methods*, Publications des séminaires de mathématiques et informatique de Rennes, 1978 (1978), pp. 1–16.

[28] J. E. Osborn, *Spectral approximation for compact operators*, Math. Comput., 29 (1975), pp. 712–725.

[29] J. M. Peña, A. LaTorre, and A. Jérusalem, *SoftFEM: The soft finite element method*, International Journal for numerical methods in engineering, 118 (2019), pp. 606–630.

[30] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical mathematics*, vol. 37 of Texts in Applied Mathematics, Springer-Verlag, Berlin, second ed., 2007.

[31] G. Strang and G. J. Fix, *An analysis of the finite element method*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1973. Prentice-Hall Series in Automatic Computation.

[32] G. M. Vainikko, *Asymptotic error bounds for projection methods in the eigenvalue problem*, Ž. Vyčisl. Mat. i Mat. Fiz., 4 (1964), pp. 405–425.

[33] ——, *Rapidity of convergence of approximation methods in eigenvalue problems*, Ž. Vyčisl. Mat. i Mat. Fiz., 7 (1967), pp. 977–987.

[34] T. Warburton and J. S. Hesthaven, *On the constants in hp-finite element trace inverse inequalities*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 2765–2773.