Contemporary Mathematics
Volume **700**, 2017
http://dx.doi.org/10.1090/conm/700/14186

Finite element methods for variational eigenvalue problems

Guido Kanschat

Contents

1. Introduction

The discretization of the eigenvalue problem through finite elements relies on two independent concepts: first, the abstract idea of a Galerkin discretization in Section 3, which describes projections to general finite dimensional subspaces of Hilbert spaces, and second, the choice of these subspaces by piecewise polynomials and the resulting concrete approximation error estimates in Section 4.

These notes are only a short overview over the theory of finite element approximation of eigenvalue problems, and to a large extend a compilation of results found in [**Bof10**, **BO87**, **BO89**]. They are also content with discussing the approximation theory and giving a pointer to a software package for experiments. After discretization, we always obtain a matrix eigenvalue problem of the form

$$Ax = \lambda Mx,$$

where A is the matrix of the discretized differential operator and M is the so-called mass matrix, which generates the bilinear form of the L^2-inner product on the finite element space. Methods for the computation of eigenvalues and eigenvectors of the discretized problem are discussed in the chapter by Chen Greif in this volume.

2010 *Mathematics Subject Classification.* Primary 65N25, 65N30.

2. Problem setting

2.1. Source and eigenvalue problems for the Laplacian. We begin by shortly reviewing the most simple differential operator in this framework and its setting in Hilbert spaces. We are beginning with the so-called *source problem*, the Dirichlet problem for Poisson's equation. Let Ω be a domain in $\mathbb{R}^d$, where the dimension d is typically 2 or 3, but not restricted to those values. The boundary $\partial\Omega$ is assumed to be Lipschitz in the sense that there is a finite covering of $\partial\Omega$ of open sets B_i and a family of Lipschitz continuous mappings Φ_i, which map the open unit ball $B \subset \mathbb{R}^d$ to Ω_i, such that $\Phi^{-1}(\Omega \cap B_i) \subset \{x \in B | x_0 > 0\}$ with x_0 being the first component of the vector x. On such a domain, we search for the solution of

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

In order to develop a reasonable solution theory for this problem, we multiply with a test function v and integrate by parts to obtain the weak formulation

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x \quad \forall v \in V, \tag{1}$$

where V is a suitable test function space, here $V = H_0^1(\Omega)$, the Sobolev space obtained by completing the space $C_0^\infty(\Omega)$ of infinitely often differentiable functions with compact support in Ω with respect to the norm

$$\|v\|_V = \sqrt{\int_\Omega |\nabla v|^2 \,\mathrm{d}x}.$$

The Riesz representation theorem guarantees that this equation has a unique solution in $u \in V$ for any suitable right hand side f, see for instance '¡[**Eva98**, **GT98**]. Indeed, while the right hand side in (1) suggests $f \in L^2(\Omega)$, we can actually relax this requirement to f being a continuous linear functional in $V' = H^{-1}(\Omega)$, such that the right hand side of (1) becomes $f(v)$. Then, the weak form redefines the Laplacian as an operator

$$-\Delta : H_0^1(\Omega) \to H^{-1}(\Omega).$$

Now, we tend to the eigenvalue problem of finding $u \in V$ and $\lambda \in \mathbb{C}$ such that

$$-\Delta u = \lambda u.$$

With the above understanding of the operator $-\Delta$, this equation does not make sense, since the object on the right is in V, the object on the left in V'. Thus, we transform this problem into weak form as well. First, we require a second space H, which in basic examples is typically chosen $L^2(\Omega)$. Then, invoke the Riesz representation theorem again, this time in H, yielding the *variational eigenvalue problem*

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \lambda \int_\Omega uv \,\mathrm{d}x \quad \forall v \in V. \tag{2}$$

Since $H_0^1(\Omega)$ is continuously embedded into $L^2(\Omega)$, the equation above is well defined.

Due to the existence and uniqueness of solutions, we can define the inverse of the Laplacian as an operator

$$(-\Delta)^{-1} : L^2(\Omega) \to H_0^1(\Omega).$$

Indeed, since the Laplacian contains two derivatives, it might be suggested that the range is contained in $H^2(\Omega)$. Nevertheless, the regularity theory of solutions on Lipschitz domains does not admit this conclusion [**Gri85**]. On the other hand, $H_0^1(\Omega)$ as the domain of the Laplacian is just what we need.

Since the embedding of $H_0^1(\Omega)$ into $L^2(\Omega)$ is not only continuous, but also compact, see for instance [**AF03**], $(-\Delta)^{-1}$ is a compact operator.

The spectral theorem for compact operators says that $(-\Delta)^{-1}$ has a point spectrum of at most countably many eigenvalues λ_k^{-1}, which may not have any accumulation point in $\mathbb{C}$ except zero. Consequently, we deduce that the spectrum of the Laplacian as an operator on $H_0^1(\Omega)$ consists of at most countably many eigenvalues $\lambda_k \neq 0$ which do not have an accumulation point in $\mathbb{C}$.

Furthermore, the Laplacian in its weak form is self-adjoint, see for instance [**GT98**], such that its spectrum is contained in $\mathbb{R}^+$. Thus, we conclude

Theorem 1. *The spectrum of the Laplacian defined as the values λ such that there is a function $u \in V$ such that equation* (2) *holds contains at most countably many real values*

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k \leq \ldots$$

2.2. Abstraction from the Laplacian. The eigenvalue problem for the Laplacian is only a model problem. Therefore, we generalize the ideas from the preceding paragraphs and introduce some notation. First, we introduce

Definition 1. A bilinear form $a(.,.)$ is bounded and elliptic on a vector space V if there are positive constants $\overline{C}$ and $\underline{C}$ such that

$$\begin{aligned} a(u,v) &\leq \overline{C}\|u\|_V\|v\|_V & \forall u,v \in V \qquad & \text{(boundedness)} \\ a(u,u) &\geq \underline{C}\|u\|_V^2 & \forall u \in V \qquad & \text{(ellipticity).} \end{aligned}$$

If in addition the bilinear form is symmetric, it defines a norm on V by

$$\|v\| = \sqrt{a(v,v)}.$$

We introduce the oprator A associated to this bilinear form defined by

$$\langle Av, w\rangle_{V^*\times V} = a(v,w) \qquad \forall v,w \in V.$$

If the form $a(.,.)$ is bounded, then $A : V \to V^*$ is a bounded operator.

Definition 2 (Source problem). Given a bilinear form $a(.,.)$ on V and f in the normed dual V^*, find a function $u \in V$ such that

$$a(u,v) = f(v) \quad \forall v \in V. \tag{4}$$

For the class of bilinear forms introduced above, the well-posedness[1] is established in the elliptic case by

[1] Well-posedness in the context of numerical approximation follows the definition of Hadamard [**Had02**], namely that the solution shall exist, be unique, and shall depend continuously on the parameters.

LEMMA 1 (Lax-Milgram). *Let $a(.,.)$ be symmetric, bounded and elliptic on V. Then a unique solution $u \in V$ of equation (4) exists with the estimate*

$$\|u\|_V \le \frac{\overline{C}}{\underline{C}} \|f\|_{V^*}, \tag{5}$$

where

$$\|f\|_{V^*} = \sup_{v \in V} \frac{|f(v)|}{\|v\|_V}$$

is the operator norm of f on the space V.

In the general case, we use a version of Banach's closed range theorem which is particularly amenable to the study of numerical approximation

LEMMA 2 (Inf-sup condition). *Let $a(.,.)$ be bounded and let there be a constant $\underline{C} > 0$ such that*

$$\inf_{\substack{u \in V \\ \|u\|=1}} \sup_{\substack{v \in V \\ \|v\|=1}} |a(u,v)| \ge \underline{C}, \qquad \inf_{\substack{v \in V \\ \|v\|=1}} \sup_{\substack{u \in V \\ \|u\|=1}} |a(u,v)| \ge \underline{C}. \tag{6}$$

Then, equation (4) has a unique solution $u \in V$ admitting the estimate (5).

Obviously, Lemma 2 implies Lemma 1 for symmetric bilinear forms.

We proceed by generalizing (2) to

DEFINITION 3 (Variational eigenvalue problem). A pair $(\lambda, u) \in \mathbb{C} \times V$ with $u \neq 0$ is an eigenpair of the operator A associated with the bilinear form $a(.,.)$, if there holds

$$a(u,v) = \lambda \langle u, v \rangle \quad \forall v \in V. \tag{7}$$

Here, $\langle .,. \rangle$ is the inner product in a Hilbert space H with norm $\|.\|$ and $V \subset H$ is the domain of A.

If the bilinear form $a(.,.)$ is nonsymmetric, we also introduce the

DEFINITION 4 (Adjoint variational eigenvalue problem). A pair $(\lambda^*, u^*) \in \mathbb{C} \times V$ with $u \neq 0$ is an adjoint eigenpair of the operator A associated with the bilinear form $a(.,.)$, if there holds

$$a(v, u^*) = \lambda^* \langle v, u^* \rangle \quad \forall v \in V. \tag{8}$$

Here, $\langle .,. \rangle$ is the inner product in a Hilbert space H with norm $\|.\|$ and $V \subset H$ is the domain of A.

It is a well known fact, that the eigenvalues λ in (7) and λ^* in (8) coincide (note that u^* is on the right hand side of the sesquilinear form $\langle .,. \rangle$). Thus, we speak of the eigenfunction u and the adjoint eigenfunction u^* with eigenvalue λ.

We point out that while the source problem was solely defined on the space V, typically the domain of the operator associated with the bilinear form, this definition of an eigenvalue problem makes use of an additional space H containing V.

An immediate consequence of ellipticity is the fact that the real parts of all eigenvalues of the variational eigenvalue problem (7) are positive and bounded from below by $\underline{C}$. If additionally the bilinear form is symmetric (Hermitean in the complex case), the eigenvalues are even real and positive, and $a(.,.)$ forms an inner product on its domain.

Let $a(.,.)$ be such that either Lemma 1 or Lemma 2 holds. Then, we define a solution operator $T : H \to V$ for the source problem such that for any $w \in H$ there holds

$$a(Tw, v) = \langle w, v\rangle \quad \forall v \in V.$$

Then, by entering Tu into (7), the eigenvalue problem is equivalent to

$$\lambda Tu = u.$$

Throughout the remainder of this chapter, we impose

Assumption 1. The solution operator T is compact, that is, the embedding $V \hookrightarrow H$ is compact.

As a consequence, the spectral theorem for compact operators holds. We denote that by this assumption, we exclude bilinear forms with zero eigenvalues, a problem which can be easily fixed by adding a shift of the form $\mu \langle u, v\rangle$ to $a(u, v)$ and computing the eigenvalues $\lambda + \mu$ of the new form.

Let $a(.,.)$ be an inner product on its domain V and V compactly embedded in H. Then, Theorem 1 holds for this variational eigenvalue problem, and thus all eigenvalues are real and positive. Thus, the spectrum can be ordered and we have two important results

Theorem 2 (Minimum principle). *The eigenpairs* (λ_k, u_k) *of the symmetric variational eigenvalue problem* (7) *are recursively characterized by*

$$\begin{aligned} u_k &= \operatorname*{argmin}_{\substack{u\in V\\ a(u,u_j)=0\ \forall j<k}} \frac{a(u,u)}{\langle u,u\rangle} \\ \lambda_k &= \frac{a(u_k,u_k)}{\langle u_k,u_k\rangle} \end{aligned}$$

Moreover, we have

Theorem 3 (Minimum-maximum theorem). *The eigenvalues* λ_k *of the symmetric variational eigenvalue problem* (7) *are characterized by*

$$\lambda_k = \min_{\substack{W\subset V\\ \dim W=k}} \max_{v\in W} \frac{a(v,v)}{\langle v,v\rangle}. \tag{9}$$

This theorem not only provides a condition that a number $\lambda \in \mathbb{R}$ be an eigenvalue, it furthermore assigns an ordering to these values:

$$0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_k \le \cdots .$$

We close this section by defining the eigenspace

$$E_k = \big\{u \in V \setminus \{0\} \big| a(u, v) = \lambda_k \langle u, v\rangle \ \forall v \in V\big\},$$

2.3. Well-posedness and stability. An important concept whenever approximating the solution of a problem is well-posedness. This concept was introduced by Hadamard in [**Had02**] and cast in the famous three conditions:

Definition 5 (Well-posedness). A problem, that assigns to a set of data x a solution $y = f(x)$ is well-posed, if

(1) A solution y exists
(2) The solution y is unique for given x

(3) The solution y depends continuously on variations of the data x

We remark that the crucial condition is the third one and that it is very vague, since neither the topology for x, nor that of y is defined. These are an important part of the model used.

The necessity of the third condition stems from the fact, that we typically cannot solve the problem with exact data, because

- data may be related to a physical configuration and is measured with some error,
- the approximation of the problem involves projected or interpolated data, and
- numbers cannot be represented exactly in a computer and must be rounded.

While the third condition of well-posedness is qualitative, it is important for the judgment of the results of a computation as well as for the evaluation of solution algorithms to quantify this continuity, basically replacing it by a Lipschitz condition:

Definition 6 (Condition number). Let $y = f(x)$ be the solution of the same abstract problem depending on data x as before, then the *condition number* κ of this problem is the smallest number, possibly infinity, such that for any data x_1, x_2 in a required set there holds

$$\|f(x_1) - f(x_2)\| \le \kappa \|x_2 - x_1\|.$$

A problem is called well-conditioned, if κ is "small" and ill-conditioned, if it is not. Here, small is defined by the requirements of the application.

Remark that even in this definition, the choice of the norms is up to its user and typically mandated by the information we require from the solution y.

As discussed in [**GW76**], the eigenvalue problem is in general ill-conditioned. Let us investigate the famous example of the $n \times n$-matrix

$$\mathfrak{A}_{n,\varepsilon} = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ \varepsilon & & & & 0 \end{pmatrix},$$

which has the only eigenvalue zero if $\varepsilon = 0$. Perturbing this matrix by a tiny positive number ε, we see that the characteristic polynomial becomes $\lambda^n + \varepsilon$. Thus, the perturbed matrix $\mathfrak{A}_{n,\varepsilon}$ has n different eigenvalues $\lambda_k(\varepsilon) = \sqrt[n]{\varepsilon} e^{ik/n}$. First, we see from this that the problem "determine the multiplicity of an eigenvalue" is ill-posed, since even the tiniest perturbation of a matrix may separate a single multiple eigenvalue into multiple single eigenvalues. The problem "determine the value of the eigenvalue" is ill-posed as well, since we do not specify which one and the solution is not unique. "Determine the eigenvalue with greatest positive real part and nonnegative imaginary part" has a unique solution, and since the formula for $\lambda_k(\varepsilon)$ is continuous, is also well-posed. But, since this formula is not Lipschitz, the condition number is infinite. Indeed, let us assume there are no other errors and numbers in a computer are stored with 16 digits accuracy. Let $n = 16$ in the definition of the matrix $\mathfrak{A}_n$ above. Then, rounding of the data may perturb the solution by the value $1/10$, which is way above the accuracy of the computer.

It is important to realize that the concepts of well-posedness and conditioning above are related to the original problem, not to the particular algorithm used for their solution. In fact, the condition number above determines the condition number of the best possible algorithm for the solution of the problem. Obviously, we may also have devised a bad algorithm. In order to judge this, we introduce as a final concept

Definition 7 (Stability of algorithms). Let $y = f(x)$ be the solution of a problem with data x and let κ be its condition number. Let $\tilde{y} = \tilde{f}(x)$ be an approximation to this solution by an algorithm $\tilde{f}$, and let $\tilde{\kappa}$ be the condition number of this problem. Then, we call an algorithm *stable*, if the constant c in $\tilde{\kappa} \le c\kappa$ is of moderate size.

Since no algorithm can have a better condition number than the original problem, we can consider the numerical solution of the example above as not computable. In fact, Golub and Wilkinson suggest in [**GW76**] to reevaluate the emphasis the Jordan normal form receives for this very reason.

Wilkinson showed [**GW76**, **Wil72**] on the other hand the following remarkable result for matrices

Theorem 4. *Let $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{n\times n}$ with $\|B\|_2 = 1$. Let λ be a simple eigenvalue of A with right eigenvector x and left eigenvector y, both of norm unity. Then, the corresponding eigenvalue $\lambda(\varepsilon)$ of the perturbed matrix $A + \varepsilon B$ admits the estimate*

$$\left|\frac{d}{d\varepsilon}\lambda(\varepsilon)\right| \le \frac{1}{y^H x}.$$

Thus, at least for very small perturbations, the conditioning of the problem of finding this eigenvalue λ depends on the angle between the right and left eigenvector. And, as an immediate corollary, we obtain that the eigenvalue problem for any simple eigenvalue of a normal matrix is well-posed, since for such matrices right and left eigenvectors coincide.

3. Galerkin approximation

The beauty and the power of Galerkin[2] methods lies in their simplicity: replace the infinite dimensional space V in which the weak formulation of differential equation problem is posed by a finite dimensional subspace, say[3] V_n. Then, we define the discrete analogues to the source and eigenvalue problems in definitions 2 and 3, respectively.

Definition 8 (Discrete source problem). Find $u_n \in V_n$ such that there holds

$$a(u_n, v_n) = \langle f, v_n\rangle \quad \forall v_n \in V_n. \tag{10}$$

Definition 9 (Discrete variational eigenvalue problem). A pair $(\lambda_n, u_n) \in \mathbb{C} \times V_n$ is a discrete eigenpair to the operator A associated with the bilinear form $a(.,.)$ if there holds

$$a(u_n, v_n) = \lambda_n \langle u_n, v_n\rangle \quad \forall v_n \in V_n. \tag{11}$$

[2] Boris Grigoryevich Galerkin (Борис Григорьевич Галёркин), 1871-1945, Russian engineer

[3] We will use the subscript n as an indicator of discretized values without giving it a defined meaning. Intuitively, it may be seen as the dimension of the finite dimensional space, but later on this will not be sufficient.

The discrete adjoint eigenvalue problem consists of finding pairs $\lambda_n^*, u_n^* \in \mathbb{C} \times V_n$ such that there holds

$$a(v_n, u_n^*) = \lambda_n^* \langle v_n, u_n^* \rangle \quad \forall v_n \in V_n.$$

Since the space V_n is finite dimensional of dimension n, we can choose a basis $\{\varphi_j\}_{j=1,\dots,n}$ and expand the solution u_n with respect to the basis functions as

$$u_n(x) = \sum_{j=1}^{n} \mathfrak{u}_j \varphi_j(x).$$

Thus, testing (11) with each φ_j, we obtain the rows of the algebraic eigenvalue problem

$$\mathfrak{A}\mathfrak{u} = \lambda_n \mathfrak{M}\mathfrak{u},$$

where we denote $\mathfrak{u} = (\mathfrak{u}_1, \dots, \mathfrak{u}_n)^T$ and the so called stiffness and mass matrices are

$$\mathfrak{A} = \Bigl(a(\varphi_j, \varphi_i)\Bigr)_{i,j=1,\dots,n}, \quad \mathfrak{M} = \Bigl(\langle \varphi_j, \varphi_i \rangle\Bigr)_{i,j=1,\dots,n}.$$

For numerical solution methods of this problem, we refer to the chapter by Chen Greif in this volume. Here, we only remind the reader of the:

THEOREM 5 (spectral theorem for normal matrices). *Let $\mathfrak{A} \in \mathbb{C}^{n \times n}$ be a normal matrix. Then, $\mathfrak{A}$ is diagonalizable and there exists an orthogonal matrix $\mathfrak{X} \in \mathbb{C}^{n \times n}$, such that*

$$\mathfrak{A} = \mathfrak{X} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \mathfrak{X}^T.$$

The column vectors of $\mathfrak{X}$ are the eigenvectors of $\mathfrak{A}$ and left and right eigenvectors for each eigenvalue coincide.

If in addition, $\mathfrak{A}$ is Hermitean, all eigenvalues are real.

3.1. The source problem. We begin our study of the properties of Galerkin approximations with the source problem and summarize analysis found in basic courses on finite element methods, for instance early chapters in [**Bra97**, **BS02**, **Cia78**, **GRS07**, **SF73**].

LEMMA 3 (Discrete Lax-Milgram). *Let $a(.,.)$ be symmetric, bounded and elliptic on V. Then a unique solution $u_n \in V_n$ of equation* (10) *exists with the estimate* (5).

PROOF. If the bilinear form $a(.,.)$ is elliptic, the Lax-Milgram Lemma 1 transfers to V_n immediately, since $V_n \subset V$. □

In the nonsymmetric case we have to assume that Lemma 2 holds in V_n as well:

LEMMA 4 (Discrete inf-sup condition). *Let $a(.,.)$ be bounded and let there be a constant $\underline{C} > 0$ such that for any n there holds*

$$\inf_{\substack{u_n \in V_n \\ \|u_n\|=1}} \sup_{\substack{v_n \in V_n \\ \|v_n\|=1}} |a(u_n, v_n)| \ge \underline{C}, \qquad \inf_{\substack{v_n \in V_n \\ \|v_n\|=1}} \sup_{\substack{u_n \in V_n \\ \|u_n\|=1}} |a(u_n, v_n)| \ge \underline{C}. \tag{12}$$

Then, equation (10) *has a unique solution $u_n \in V_n$ admitting the estimate* (5).

In principle, the constants $\underline{C}$ and $\overline{C}$ in Lemmas 2 and 4, respectively may be different. By taking the worse pair, we can keep the presentation simple and consider them equal. Indeed, Lemma4 would even hold with constants depending on n. We do not consider this case because it leads to suboptimal convergence estimates.

We assume unique solvability of the discrete problem either by Lemma 3 or by Lemma 4. Thus, let $u \in V$ and $u_n \in V_n$ be the unique solutions of

$$\begin{aligned} a(u,v) &= f(v) && \forall v \in V \\ a(u_n,v_n) &= f(v_n) && \forall v_n \in V_n. \end{aligned}$$

An immediate consequence obtained by testing with $v \in V_n \subset V$ in the first equation and subtracting the second is

Lemma 5 (Galerkin orthogonality). *There holds*

$$a(u-u_n,v_n) = 0 \quad \forall v \in V_n.$$

From this result, we immediately obtain

Lemma 6 (Céa's lemma). *Let $a(.,.)$ be bounded and elliptic such that equations* (3) *hold. Let furthermore u and u_n be the solutions in equations* (13)*. Then, there holds the quasi-best approximation result*

$$\|u-u_n\|_V \le \frac{\overline{C}}{\underline{C}} \inf_{v\in V_n} \|u-v\|_V.$$

Proof. From the assumptions there holds for arbitrary $w_n \in V_n$:

$$\begin{aligned} \frac{1}{\underline{C}}\|u-u_n\|_V^2 &\le a(u-u_n,u-u_n) && \text{ellipticity} \\ &= a(u-u_n,u-w_n) && \text{Galerkin orthogonality} \\ &\le \overline{C}\|u-u_n\|_V\,\|u-w_n\|_V && \text{boundedness.} \end{aligned}$$

Dividing by $\|u-u_n\|_V$ and making use of the fact that w_n was chosen arbitrarily in V_n yields the lemma. □

Thus, the error between the continuous solution $u \in V$ and the discrete solution $u_n \in V_n$ is reduced to the approximability of the function $u \in V$ by a function in V_n. This is now simply a property of the approximation space and independent of the bilinear form $a(.,.)$. Thus, we introduce the notation

$$\overline{\eta}_u(V_n) = \inf_{v\in V_n} \|u-v\|_V,$$

which describes the approximation accuracy of u by elements of V_n. The estimate in Céa's lemma thus becomes

$$\|u-u_n\|_V \le \frac{\overline{C}}{\underline{C}}\overline{\eta}_u(V_n).$$

3.2. The discrete eigenvalue problem. The results on eigenvectors and eigenspaces in this section will all be of the same type. The choice of finite element spaces for V_n and their approximation properties are deferred to Section 4. Here, we assume that the sequence of spaces $\{V_n\}$ is exhausting V, that is,

$$\forall u \in V \ \inf_{v\in V_n} \|u-v\|_V \to 0 \qquad \text{as } n\to\infty. \tag{14}$$

We introduce two operators important for the analysis of eigenvalue problems. First, the Ritz projection $P_n : V \to V_n$ is defined such that for any $w \in V$ there holds

$$a(P_n w, v_n) = a(w, v_n) \quad \forall v_n \in V_n.$$

In addition, we define the discrete solution operator T_n, such that for any $w_n \in V_n$ there holds

$$a(T_n w_n, v_n) = (w_n, v_n) \quad \forall v_n \in V_n.$$

Let P_n^* be the $a(.,.)$-adjoint of P_n, $P_n = P_n^*$ for a symmetric bilinear form. For any $v \in V$ there holds

$$a(T_n w_n, v) = a(T_n w_n, P_n^* v) = (w_n, P_n^* v) = a(T w_n, P_n^* v) = a(P_n w_n, v).$$

Thus, we obtain the relation

$$T_n = P_n T.$$

Entering $T_n u_n$ into (11), we obtain the equivalent eigenvalue problem

$$\lambda_n P_n T u_n = u_n.$$

In this context, the assumption that the sequence of discrete spaces exhausts V in (14) translates into

$$\|T - P_n T\| = \sup_{\|u\|_V = 1} \|Tu - P_n T u\|_V \to 0 \qquad \text{as } n \to \infty. \tag{15}$$

3.3. A single eigenpair. In this section, we derive error estimates for one solution of equation (11) in comparison to those of equation (7). We first follow [**Fix73**] for the general, non-selfadjoint problem[4]. We assume that (6) holds for the space V as well as for the finite dimensional subspace V_n with the same constant $\underline{C}$ independent of n. For an eigenvalue λ of the continuous problem (7), its eigenspace E_λ, and its adjoint eigenspace E_λ^*, we characterize approximation by V_n by

$$\overline{\eta}_\lambda(V_n) = \sup_{\substack{u \in E_\lambda \\ \|u\|=1}} \inf_{v \in V_n} \|u - v\|_V,$$
$$\overline{\eta}_\lambda^*(V_n) = \sup_{\substack{u^* \in E_\lambda^* \\ \|u^*\|=1}} \inf_{v \in V_n} \|u^* - v\|_V,$$

that is, we measure the best approximation of the unit vector worst approximated in this eigenspace. From [**Fix73**, Theorem 1], we obtain the following estimates

THEOREM 6. *Let the inf-sup condition* (6)*, the discrete inf-sup condition* (12)*, and the uniform convergence* (15) *hold. Let λ be an eigenvalue of* (7)*. Then, there is an eigenvalue λ_n of* (11)*, such that*

$$|\lambda_n - \lambda| \le C \overline{\eta}_\lambda(V_n) \overline{\eta}_\lambda^*(V_n). \tag{17}$$

Furthermore, if $u_n \in V_n$ with $\|u_n\| = 1$, there is a function $u \in E_\lambda$ such that

$$\|u_n - u\|_V \le C \overline{\eta}_\lambda(V_n). \tag{18}$$

In both estimates, the constant C depends on the bilinear forms, but not on n.

[4] We only consider the case of diagonalizable operators here. See [**Fix73**] for the general case.

From estimate (18), we deduce that eigenvectors are approximated with the same accuracy as solutions to the source problem. Eigenvalues, as (17) shows, typically converge twice as fast as $n \to \infty$.

3.4. Multiple eigenvalues and their eigenspaces. The result in the previous section covers multiple eigenvalues, but it does not make any predictions how a multidimensional eigenspace is approximated. Such a theory was developed in [**BO87**, **BO89**]. It makes much stronger statements on the approximation of eigenvectors, but it assumes an order relation on the spectrum. Thus, here we assume a real symmetric bilinear form and real eigenvalues.

In the case of a multiple eigenvalue of the continuous problem, the approximating eigenvalues may be either simple or or multiple. Obviously, we have the

THEOREM 7 (Discrete min-max condition).

$$\lambda_{n,k} = \min_{\substack{W\subset V_n\\ \dim E=k}} \max_{v\in W} \frac{a(v,v)}{\langle v,v\rangle}. \tag{19}$$

Since the maximum in this equation only depends on the choice of the space W and the minimum is taken over a smaller set than in equation (9), we always have $\lambda_{k,n} \ge \lambda_k$.

We now consider the case of a q-fold eigenvalue of the continuous problem, that is,

$$\lambda_{k-1} < \lambda_k = \dots = \lambda_{k+q-1} < \lambda_{k+q}.$$

Due to the operator approximation (15), each discrete eigenvector converges to its continuous counterpart. Thus, if n is sufficiently large, we have the situation

$$\lambda_k \le \lambda_{n,k} \le \dots \le \lambda_{n,k+q-1} < \lambda_{k+q}.$$

We have (see e.g. [**BO89**]) the following quasi-orthogonality condition, which relates the error of the eigenvalue to the residual of the eigenvalue problem:

LEMMA 7. *Assume that* (λ, u) *is an eigenpair according to Definition* 3 *and* $w \in V$ *arbitrary with* $\|w\|_H = 1$*. Furthermore, let*

$$\mu = a(w,w).$$

Then,

$$\mu - \lambda = a(w-u, w-u) - \lambda\|w-u\|^2.$$

PROOF. From the variational eigenvalue problem, we obtain $a(u,u) = \lambda\|u\|^2$ and by symmetry

$$a(v,u) = \lambda \langle v,u\rangle \quad \forall v \in V.$$

Thus, entering the assumptions, we obtain

$$\begin{aligned} a(w-u,w-u) - \lambda\|w-u\|^2 =& a(w,w) - 2a(w,u) + a(u,u) \\ & - \lambda\|w\|^2 + 2\lambda\langle w,u\rangle - \lambda\|u\|^2 \\ =&\mu\|w\|^2 + \lambda\|u\|^2 - \lambda\|w\|_2 - \lambda\|u\|^2 \\ =&\mu - \lambda. \end{aligned}$$

□

Next we refine the notion of approximability of eigenvectors. In (15) and consequently in (17) and (18), only the worst approximation within the eigenspace E_λ was used in the estimate. Here, we define recursively and begin with

$$\eta_{k,0}(V_n) = \inf_{\substack{u\in E_{\lambda_k}\\ \|u\|=1}} \inf_{v\in V_n} \|u-v\|_V .$$

Let u_0 be a vector for which this infimum is achieved, and similarly u_j below. Define for $j = 1,\dots,q-1$

$$E^j_{\lambda_k} = \bigl\{u \in E_{\lambda_k} \bigm| \langle u, u_i\rangle = 0,\ i = 0,\dots,j-1\bigr\}.$$

Then, define

$$\eta_{k,j}(V_n) = \inf_{\substack{u\in E^j_{\lambda_k}\\ \|u\|=1}} \inf_{v\in V_n} \|u-v\|_V .$$

Clearly, $\eta_{k,j}(V_n) \le \overline{\eta}_{\lambda_k}(V_n)$, such that estimates involving these quantities are sharper than the previous ones. We obtain the estimate [**BO87**, Theorem 3.1]:

THEOREM 8. *For every eigenvalue λ_k of* (9) *of multiplicity q, there is an n_0 and a constant C such that for any $n \ge n_0$ and any discrete eigenvalue $\lambda_{n,k+j}$ of* (19) *with $0 \le j < q$ there holds*

$$\lambda_{n,k+j} - \lambda_k \le C\eta^2_{k,j}(V_n).$$

Furthermore, for each discrete eigenfunction $u_{n,k+j}$ solving (11) *for $\lambda_{n,k+j}$, there is a continuous eigenfunction u_{k+j}, such that*

$$\|u_{n,k+j} - u_{k+j}\|_V \le C\eta_{k,j}(V_n).$$

First, this theorem shows that the eigenvalue estimate in the previous section may be pessimistic and it should be possible to obtain an error bounded by $\eta^2_{\lambda_k,0}$ instead of $\overline{\eta}^2_{\lambda_k}$. But the consequences of this theorem are reaching farther. It predicts, that the kth discrete eigenvalue approximates the kth continuous eigenvalue, therefore preserving the structure of the spectrum. Furthermore, in the case that a continuous multiple eigenvalue is "split" into different discrete eigenvalues, the order of these is determined by the approximability of the corresponding eigenfunction. Thus, the method will in fact compute approximations for the functions $u_0,\dots,u_{q-1}$ in the definition of $\eta_{k,j}$ above in exactly this order. See Problem 5 in Section B.1.

4. The finite element method

In this section, we give a cursory introduction to finite element spaces and their approximation properties. We will not present the method in all detail and refer the reader to the textbooks [**Bra97**, **BS02**, **Cia78**, **GRS07**, **SF73**].

The fundamental property of the finite element method as a Galerkin scheme consists in defining a discrete subspace for approximation. Thus, in this section we construct the spaces and provide estimates for the right hand sides of the quasi best approximation results of the previous section.

4.1. Finite element spaces. Finite element spaces consist of three components:

(1) A mesh consisting of mesh cells
(2) A shape function space for each mesh cell
(3) Node functionals

4.1.1. *Meshes and mesh cells.* A mesh[5] T_h is a covering[6] of the computational domain Ω by nonoverlapping cells of simple geometry. Here, simple typically means simplices[7] or smooth images of hypercubes in any dimension. We assume that for each topology type of a mesh cell T there is a reference cell $\hat{T}$. For simplices, this might be the simplex with the origin and the unit vectors e_i as its corners. For hypercubes, this might be the reference hypercube $[0,1]^d$. For each cell of the mesh T_h, there is a one-to-one mapping $\Phi : \hat{T} \to T$. The inverse of this mapping can be decomposed into the following parts:

(1) A fixed body movement Φ_F^{-1} (translation and rotation) such that the longest edge of T is a subset of the positive x_1-axis and the barycenter has only positive coordinates.
(2) An isotropic scaling Φ_S^{-1} by the scaling factor $1/h_T$ such that this longest edge is mapped to $[0,1]$.
(3) A warping operation Φ_W^{-1}.

We compose

$$\Phi = \Phi_F \circ \Phi_S \circ \Phi_W, \qquad \nabla\Phi = \nabla\Phi_F \nabla\Phi_S \nabla\Phi_W.$$

By construction, $\nabla\Phi_F$ is orthogonal and $\Phi_S = h_T\mathbb{I}$. The "shape" of T is encoded in Φ_W. We call a family of meshes $\{T_h\}$ shape regular, if there is a constant C such that for any cell T of any of the meshes we can find such a decomposition such that for the singular values $\sigma_i(\hat{x})$ of $\Phi_W(\hat{x})$ holds

$$\max_{\hat{x}\in\hat{T}} \max_{i=1,\dots,d} \sigma_i(\hat{x}) \le C \min_{\hat{x}\in\hat{T}} \min_{i=1,\dots,d} \sigma_i(\hat{x}).$$

Several sufficient geometric conditions for this have been introduced in the literature, namely for simplices, that the circumference is bounded uniformly by the radius of the inscribed sphere. For quadrilaterals, all angles should be uniformly bounded away from π and two vertices should not get too close to each other.

4.1.2. *Shape functions.* Shape functions form the local function spaces on each mesh cell. In almost all finite element methods, they are polynomial spaces or derived from such. It is convenient to derive them on the reference cell $\hat{T}$ and define them on the mesh cell T by pull-back through the mapping Φ. Standard spaces are the space P_k of multivariate polynomials of degree k, namely (for example in three dimensions)

$$P_k = \operatorname{span}\bigl\{x^\alpha y^\beta z^\gamma \bigm| 0 \le \alpha,\beta,\gamma \;\wedge\; \alpha+\beta+\gamma \le k\bigr\},$$

[5] The index h is used to be able to denote sequences of meshes. It is loosely understood to mean mesh size. On the other hand, modern finite element methods do not use a fixed mesh size, such that the meaning of h is somewhat diffuse.

[6] A favorite discussion in finite element expositions is whether cells should be open or closed. When we consider a covering or the intersection of cells, they should be closed. When we write nonoverlapping, they should be open. In this text we decide to be undecided, they may be open or closed, whatever is befitting at the moment. The context and a second of thought will provide the right meaning.

[7] deal.II and thus Amandus do not provide simplicial meshes. Other software does.

and the space of tensor product polynomials of degree k, namely

$$Q_k = \operatorname{span}\{x^\alpha y^\beta z^\gamma \big| 0 \le \alpha, \beta, \gamma \le k\}.$$

4.1.3. *Node functionals.* Node functionals establish continuity over cell boundaries. For instance, in order that a piecewise polynomial function be in the space $H^1(\Omega)$, it must be continuous. This can be achieved by the following mechanism: if the mesh cell is a cube, the trace of a polynomial in Q_k on one of its faces is again in Q_k, just of one dimension lower. The trace on one of its edges is again Q_k, but only in dimension one. Thus, in order to establish continuity of functions between sharing a face, it is sufficient to establish interpolation conditions on the face and require that they are equal for the functions on both cells. In order to achieve continuity, we have to do this for vertices, edges, and faces. A very instructive graphical representation can be found in [**AL14**].

Node functionals define the continuity between cells, but they also determine the topology of the global (in the sense of "on the whole mesh") finite element space V_n. In particular, the number of node functionals scattered over the mesh is the dimension of the finite element space.

4.1.4. *Approximation properties.* Error estimates for finite element functions can be derived either by averaged Taylor expansion of Sobolev functions [**BS02**] or by an abstract argument [**Cia78**]. They use an interpolation operator based on node functionals and the degree of the shape function spaces to

LEMMA 8. *Let $u \in H^s(\Omega)$, and let $\mathbb{T}_n$ be a mesh of mesh size $h = \max h_T$. Let V_n be a finite element space on $\mathbb{T}_n$ based on a shape function space containing P_k. Then, there is a function $v_n \in V_n$, such that*

$$|u - v_n|_{H^m} \le C h^{\min s-1, k+1-m} |u|_{H^s}.$$

From this estimate follows immediately that the quantities $\overline{\eta}_\lambda(V_n)$ and $\eta_{k,j}(V_n)$ defined in (16) and (20), respectively, are of order k if the eigenfunctions are sufficiently smooth. On the other hand, since eigenfunctions become more and more oscillatory for larger eigenvalues, this estimate also implies, that these values grow for higher eigenvalues. The largest eigenvalues of the discrete problem correspond to functions oscillating with a period corresponding to the grid spacing, and are thus determined more by mesh geometry and shape function spaces than by the shape of continuous eigenfunctions. Thus, only smaller eigenvalues and their eigenfunctions are approximated reliably.

5. Saddle point problems

So far we have only considered elliptic problems. In this section, we are going to extend our theory to problems with constraints. The two model problems we consider with increasing difficulty are the Stokes and the Maxwell eigenvalue problem. In both cases, the mathematical formulation is based on Lagrange multipliers and leads to saddle point problems. We provide their basic theory first and then study the two applications. The material in this section can be found for instance in [**BFB13**].

5.1. Saddle point problems. Consider a symmetric, positive definite bilinear form $a(.,.)$ on the space V. Consider a second bilinear form $b(v, q)$ with the arguments $v \in V$ and q from a second space Q. Now, we formulate the minimization problem: find $u \in V$, such that

$$a(u,u) = \min_{v\in V}\bigl[a(v,v) - \langle f,v\rangle_V\bigr], \qquad \text{subject to } b(v,q) = \langle g,q\rangle_Q\,, \quad \forall q \in Q,$$

that is, the bilinear form $b(.,.)$ constrains the original minimization problem in V. The Lagrange multiplier rule says, that the solution u of this constrained minimization problem is a stationary point of the Lagrange functional

$$\mathcal{L}(u,p) = \bigl[a(u,u) - \langle f,u\rangle_V\bigr] - \bigl[b(u,p) - \langle g,p\rangle_Q\bigr].$$

Closer analysis reveals that locally it is the minimum with respect to variations of u and the maximum with respect to variations of p, hence the term saddle point problem. The Euler-Lagrange equations of this system can be written as: find $(u,p) \in V \times Q$ such that for all $v \in V$ and $q \in Q$ there holds

$$a(u,v) - b(v,p) - b(u,q) = \langle f,v\rangle_V + \langle g,q\rangle_Q\,. \tag{21}$$

Here, we already have the immediate analogue of the weak formulation (6), the source problem for a saddle point formulation. It is customary to consider (21) more in the form of a system of equations for the vector $(u,p)^T$, yielding the equivalent formulation

$$\begin{array}{lll} a(u,v) - b(v,p) & = \langle f,v\rangle_V & \forall v \in V \\ b(u,q) & = \langle g,q\rangle_Q & \forall q \in Q. \end{array} \tag{22}$$

In order to simplify the presentation, we will assume $g = 0$, which is also the physically relevant situation in both our examples.

Since we have already established that the solution is a saddle point, the system must be indefinite and therefore cannot be elliptic. Thus, well-posedness must be derived by the inf-sup condition. For saddle point problems, this condition takes a special form. To this end, let us first define the space of functions which obey the constraint, namely

$$V^0 = \bigl\{v \in V \big| b(v,q) = 0\ \forall q \in Q\bigr\}.$$

Then, u is a solution of either (21) or (22) if and only if it is a solution to the reduced problem: find $u \in V^0$ such that

$$a(u,v) = \langle f,v\rangle \qquad \forall v \in V^0. \tag{23}$$

Thus, we can derive unique existence of u without considering p. In a second step, we can determine p from the equation

$$b(v,p) = \langle f,v\rangle_V - a(u,v) \qquad \forall v \in V.$$

Thus, we need an inf-sup condition for u on the subspace V^0 and one for p with respect to the form $b(.,.)$.

LEMMA 9 (Mixed inf-sup condition). *Let $a(.,.)$ be bounded on $V \times V$ and $b(.,.)$ be bounded on $V \times Q$. Let $a(.,.)$ admit the inf-sup condition* (6) *on the space V^0. Assume furthermore that for $b(.,.)$ there holds*

$$\inf_{\substack{q\in Q\\ \|u\|=1}} \sup_{\substack{v\in V\\ \|v\|=1}} |b(v,q)| \ge \underline{B}. \tag{24}$$

Then, there exists a unique solution $(u,p) \in V \times Q$ of equations (21) *and* (22), *respectively.*

Considering eigenvalue problems of equations in saddle point form we will adopt the variational setting of Section 2. But, here we are faced with two options, depending on whether we consider eigenvalues of the whole operator on the left of (22) with associated eigenvalues in $V \times Q$ or eigenvalues of the constrained problem (23). Accordingly, we define

Definition 10 (Eigenvalue problem for the saddle point operator). A triplet $(\lambda, u, p) \in \mathbb{C} \times V \times Q$ with $(u, p) \neq 0$ is an eigenpair of the operator associated with the saddle point problem (22), if there holds

$$a(u,v) - b(v,p) - b(u,q) = \lambda\big[\langle u,v\rangle_H + \langle p,q\rangle_Q\big], \qquad \forall v \in V, q \in Q. \tag{25}$$

Definition 11 (Constrained eigenvalue problem). A pair $(\lambda, u) \in \mathbb{C} \times V$ with $u \neq 0$ is an eigenpair of the operator A associated with the bilinear form $a(.,.)$ on the subspace V^0, if there is a function $p \in Q$ such that there holds

$$a(u,v) - b(v,p) - b(u,q) = \lambda\,\langle u,v\rangle_H\,, \qquad \forall v \in V, q \in Q. \tag{26}$$

Note that in the first definition $(u,p) \in V \times Q$ is considered the eigenfunction, while in the second definition, $u \in V$ is the eigenfunction and $p \in Q$ is just the Lagrange multiplier for the constraint. Note also that the two variational formulations only differ by the inner products on the right hand side.

5.2. The Stokes equations. The Stokes equations for incompressible flow are formulated for a vector valued velocity u and a scalar pressure. We choose the simplest setting here and refer to [**BFB13**] for more details: the boundary condition is no-slip, that is, all velocities vanish at the boundary of the domain. Thus, the velocity space is $V = H^1_0(\Omega;\mathbb{R}^d)$. The matching pressure space in the sense of (24) is

$$Q = \left\{ q \in L^2(\Omega) \middle| \int_\Omega q\,\mathrm{d}x = 0 \right\}.$$

The incompressibility constraint is $\nabla \cdot u = 0$, such that we choose

$$b(v,q) = \int_\Omega q \nabla \cdot v\,\mathrm{d}x.$$

Finally, the bilinear form $a(.,.)$ of our abstract framework has to be defined. While the form should involve the strain tensor for reasons of frame invariance, it turns out that no-slip boundary conditions allow for a simpler form: let u_i be the components of the velocity vector, then

$$a(u,v) = \sum_{i=1}^d \int_\Omega \nabla u_i \cdot \nabla v_i\,\mathrm{d}x.$$

The next step involves finding discrete spaces. Like in Section 2, we have to require that the inf-sup condition holds in the discrete setting. The operator $a(.,.)$ is elliptic on V, thus, the inf-sup condition (6) holds for any subspace. It remains to guarantee (24). In order to find conditions on the spaces V_n and Q_n, we rephrase (24) to

$$\forall q \in Q_h\; \exists v \in V_h\; : \|v\|_V = \|q\|_Q \;\wedge\; b(v,q) \geq \underline{B}\|q\|_Q^2.$$

In this form, we see that (24) is among others a condition that the velocity is "big enough" to control the pressure. Many pairs of spaces have been proposed over the last decades. Here, we point out the Hood/Taylor pair Q_{k+1}/Q_k and the divergence conforming discontinuous Galerkin method, see [**CKS07**,**KS14**]. Both of them are implemented in the programs described in Section A.

From the subspace property $V^0 \subset V$, it is clear that constrained eigenvalues of an elliptic bilinear form are greater than unconstrained ones. Otherwise, after obeying the mixed inf-sup condition, the Stokes problem does not pose any more challenges than the problems we considered before. In particular, we do not have to expect spurious eigenvalues inserted into the spectrum and the theory of subsection 3.4 holds.

5.3. The Maxwell equations. Maxwell's equations have a similar setup as Stokes equations in the way that they also compute a vector field, here the magnetic field, under a divergence constraint. But, the bilinear form $a(.,.)$ is changed to one based on the curl of the vector field.

$$a(u,v) = \int_\Omega \nabla \times u \cdot \nabla \times v \,\mathrm{d}x.$$

The natural space for solutions is the graph space

$$V = H_0^{\mathrm{curl}}(\Omega) = \left\{ v \in L^2(\Omega;\mathbb{R}^d) \middle| \nabla \times v \in L^2(\Omega;\mathbb{R}^d) \ \wedge v \times n = 0 \text{ on } \partial\Omega \right\}.$$

The difficulties arising are two-fold: first, the divergence of a vector field in $H_0^{\mathrm{curl}}(\Omega)$ is not well-defined. Thus, we have to integrate the divergence condition by parts to obtain

$$b(v,q) = \int_\Omega v \cdot \nabla q \,\mathrm{d}x. \tag{27}$$

The natural space for the so-called pseudo pressure p is thus $Q = H_0^1(\Omega)$. The second difficulty results from the fact that the curl operator has a big kernel, namely all gradients, plus additional functions if the domain is not simply connected. For a thorough study of the involved topics of cochain complexes of Hilbert- and finite element spaces, we refer the reader to [**AFW06**, **AFW10**], where the analytical framework is derived. Here, we just note that we have the Hodge decomposition of $H_0^{\mathrm{curl}}(\Omega)$ into

$$V = H_0^{\mathrm{curl}}(\Omega) = \underbrace{\nabla Q \oplus \mathcal{H}}_{=\ker \nabla\times} \oplus V^\perp.$$

Here, $\mathcal{H}$ are the harmonic forms, which are computed in Problem 4 in Appendix B. This decomposition is in fact orthogonal in $L^2(\Omega;\mathbb{R}^d)$, and the bilinear form $a(.,.)$ is elliptic just on $V^\perp$.

We surely do not want to compute a basis for the kernel of the curl operator. What we are interested in is the dimension of $\mathcal{H}$, since it is equal to the Betti numbers and determines the topology of the domain. Furthermore, we are interested in the nonzero eigenvalues when we constrain the problem to $V^\perp$. Thus, we are in the framework of the constrained eigenvalue problem (26) and the space V_0 is defined by the form $b(.,.)$ in equation (27).

When we investigated the Stokes problem, we concluded that the discrete velocity space had to be big enough to control the discrete pressure. Here now, we also need the opposite mechanism: the discrete space for the pseudo pressure must be big enough to guarantee that for any $v \in V_h$ there holds

$$b(v,q) = 0 \;\forall q \in Q_h \quad \Rightarrow \quad b(v,q) = 0 \;\forall q \in Q. \tag{28}$$

Thus, the spaces V_h and Q_h must match exactly, which can be achieved by the mechanisms in [**AFW06**, **AFW10**].

But, does it actually matter? The results of problem 3 in Appendix B.2 indicate that the spectrum may be completely destroyed and spurious eigenvalues may be inserted at any point of the spectrum. How can this happen?

Let $u_n \in V_n$ be an eigenfunction with eigenvalue

$$\lambda_n = \frac{a(u_n, u_n)}{\langle u_n, u_n \rangle},$$

such that the left hand side of (28) holds, but not the right. Then, we can split orthogonally $u_n = u_n^0 + u_n^\perp$, where the right hand side of (28) holds for u_n^0. Thus

$$\lambda_n = \frac{a(u_n^0, u_n^0)}{\langle u_n^0, u_n^0 \rangle + \langle u_n^\perp, u_n^\perp \rangle}.$$

We see that even a high frequency function can produce a small eigenfunction, if its part in V^0 is sufficiently small.

Appendix A. Programs for experiments

Basic experiments with the finite element method for solving source and aigenvalue problems can be conducted with our sofware package Amandus. Amandus in turn is based on the finite element library deal.II available at `www.dealii.org`.

Note that the instructions below require a Unix system like Linux or MAC OSX. Neither deal.II nor Amandus work reliably on native Windows as of now. If you only have a Windows computer available, either use a virtual box or consider a dual boot installation with Linux.

A.1. The Amandus program package. Amandus is a collection of a few classes which simplify the usage of the quite complex finite element libary deal.II combined with a set of example applications. It can be cloned from the git archive at `https://bitbucket.org/guidokanschat/amandus`.

A.1.1. *Installation.* First, you have to install a recent version of deal.II. When configuring with cmake, make sure that the Arpack library is picked up. The configuration output lists all configured options at the end. if you see a line

```
DEAL_II_WITH_ARPACK set up with external dependencies
```

you have it. If you see instead

```
( DEAL_II_WITH_ARPACK = OFF )
```

check your system installation. For instance, on Debian and related system, you have to install the developer package of libarpack as well.

cmake and deal.II use three different directory: the source directory which you downloaded, the build directory in which you configure and build the library, and an install directory into which you install the library. A reasonable layout for these might be

```
$HOME/dealii/deal.II (source)
$HOME/dealii/build
$HOME/dealii/install
```

How do we get there? First, in your home directory, make a subdirectory `dealii`. Change into this directory, create a subdirectory `build`, and unpack deal.II, either the download of a release or the clone from github into the subdirectory `deal.II`. Change to the directory `build` created before and run cmake ../deal.II.

`Note:` Amandus is currently under co-development with some functionality of deal.II. Therefore, we recommend to clone the developer version of deal.II (see the deal.II web site for this).

You are free to delete the build directory after make install, and even the source directory, if you are not planning to rebuild deal.II.

Installation of Amandus follows the same concept, except that you might skip the installation and run code in the build directory right away. Given the above directories for deal.II and the structure

```
$HOME/amandus/amandus (source)
$HOME/amandus/build
```

for Amandus, you can configure Amandus in the build directory with

```
cmake -DDEAL_II_DIR=$HOME/dealii/install ../amandus
```

In order to try if everything is installed correctly, do in the build directory

```
make laplace_eigen
./laplace/eigen
```

You should get approximations of the eigenvalues of the Laplacian on the square $[-1,1]^2$.

A.1.2. *Experimenting with eigenvalue problems.* The Amandus source directory contains subdirectories for different partial differential equations and some of them contain code for eigenvalue problems. Check for files `eigen.h` and `eigen.cc`. By the time of writing these notes, they exist in the subdirectories `laplace`, `stokes`, and `maxwell`. Replace `XXX` below for any of these. All of these directories contain a file `eigen.prm`, which is copied by cmake into the corresponding subdirectory of the build directory.

In the file `amandus/build/XXX/eigen.prm`, change the number of eigenvalues or the number of refinement iterations (`Steps`). Use the output section to output eigenfunctions in gnuplot or VTK format; visualize them with gnuplot or a VTK viewer (check out paraview or visit). Experiment with different finite elements (`FE`). All these can be tried without recompiling the program.

Try different domains and coarse meshes. In order to do this, you must change the file `amandus/amandus/XXX/eigen.cc` and recompile afterwards. Compilation can always be done in the build directory with

```
make XXX_eigen
```

Find the line which contains the word GridGenerator and change the function, for instance to GridGenerator::simplex. More ideas to get started are in Appendix B.

Appendix B. Selected problems

B.1. The Laplacian. As a basis for these problems, use the programs in `amandus/laplace` referring to eigenvalue problems, in particular `eigen.cc`. See Appendix A for compiling, running, and modifying them.

(1) Verify the theoretical convergence estimates by computing on a sequence of meshes, where each mesh is obtained by global refinement of the previous one.

(a) If the exact eigenvalue λ is known, by computing the error $\lambda_h - \lambda$ on each mesh. Use the assumption

$$\lambda_h - \lambda = Ch^p$$

on consecutively refined meshes to estimate C and p. How do they depend on the mesh size and the eigenvalue?

(b) If the exact eigenvalue is not known, use the "intrinsic convergence rate" generated by terms of the form $\|\lambda_h - \lambda_{h/2}\|$ instead of $\lambda_h - \lambda$ and conclude with properties of the geometric series.

(2) Change the polynomial order (`build/laplace/eigen.prm`) and check how the results of the previous exercise change.

(3) Create your own program file `my.cc` in the subdirectory `laplace` and run `cmake .` in the build directory to add it to the compilation list.

(4) Use GridTools::distort_random() to break the symmetry of the mesh and see how multiple eigenvalues separate.

(5) The code in `eigen.cc` uses regular divisions of a mesh for a square with a single cell obtained by GridGenerator::hyper_cube(). Change this to an anisotropic mesh by using GridGenerator::subdivided_hyper_rectangle() and see how the approximations of eigenvalues 2 and 3 and their eigenfunctions changes.

(6) Change the domain from a square to

(a) an L-shaped and a slit domain, a triangle

(b) a circle (Hint: dealii tutorial step 6)

(c) a cube (Hint: dealii tutorial step 4) and experiment with paraview visualization of functions in 3D

(7) Introduce a variable coefficient

(a) First, select cells with a certain criterion and add a factor to the call of the Laplace::cell_matrix(). Example:

```
double factor = 1.;
if (dinfo.cell->center()(0) < 0.5 &&
    dinfo.cell->center()(0) < 0.5)
  factor = 10.;
Laplace::cell_matrix(dinfo.matrix(0,false).matrix,
  info.fe_values(0), factor);
```

(b) Find out how to use continuously varying coefficients (requires some more programming)

(8) Implement error estimation and adaptive refinement (big exercise)

B.2. Mixed problems.

(1) Change the code for the Stokes eigenvalue problem such that it solves the eigenvalue problem (25) of the saddle point operator instead of the constrained eigenvalue problem (26).

(2) Compute eigenvalues of the Oseen equation by adding an advection term (namespace LocalIntegrators::Advection) to the bilinear form $a(u, v)$.

(3) For the Maxwell eigenvalue problem, check how the eigenvalues change if you replace FE_Nedelec in the parameter file by FE_System[FE_Q(1)^d]

(4) Compute the zero eigenvalues of the Maxwell operator and of the Stokes operator on a domain with holes generated with GridGenerator::cheese() and compare.

References

[AF03] Robert A. Adams and John J. F. Fournier, *Sobolev spaces*, 2nd ed., Pure and Applied Mathematics (Amsterdam), vol. 140, Elsevier/Academic Press, Amsterdam, 2003. MR2424078

[AFW06] Douglas N. Arnold, Richard S. Falk, and Ragnar Winther, *Finite element exterior calculus, homological techniques, and applications*, Acta Numer. **15** (2006), 1–155, DOI 10.1017/S0962492906210018. MR2269741

[AFW10] Douglas N. Arnold, Richard S. Falk, and Ragnar Winther, *Finite element exterior calculus: from Hodge theory to numerical stability*, Bull. Amer. Math. Soc. (N.S.) **47** (2010), no. 2, 281–354, DOI 10.1090/S0273-0979-10-01278-4. MR2594630

[AL14] D. N. Arnold and A. Logg. Periodic table of the finite elements. *SIAM News*, 47(9), 2014. `www.femtable.org`.

[BFB13] Daniele Boffi, Franco Brezzi, and Michel Fortin, *Mixed finite element methods and applications*, Springer Series in Computational Mathematics, vol. 44, Springer, Heidelberg, 2013. MR3097958

[BO87] I. Babuška and J. E. Osborn, *Estimates for the errors in eigenvalue and eigenvector approximation by Galerkin methods, with particular attention to the case of multiple eigenvalues*, SIAM J. Numer. Anal. **24** (1987), no. 6, 1249–1276, DOI 10.1137/0724082. MR917451

[BO89] I. Babuška and J. E. Osborn, *Finite element-Galerkin approximation of the eigenvalues and eigenvectors of selfadjoint problems*, Math. Comp. **52** (1989), no. 186, 275–297, DOI 10.2307/2008468. MR962210

[Bof10] Daniele Boffi, *Finite element approximation of eigenvalue problems*, Acta Numer. **19** (2010), 1–120, DOI 10.1017/S0962492910000012. MR2652780

[Bra97] Dietrich Braess, *Finite elements*, Cambridge University Press, Cambridge, 1997. Theory, fast solvers, and applications in solid mechanics; Translated from the 1992 German original by Larry L. Schumaker. MR1463151

[BS02] Susanne C. Brenner and L. Ridgway Scott, *The mathematical theory of finite element methods*, 2nd ed., Texts in Applied Mathematics, vol. 15, Springer-Verlag, New York, 2002. MR1894376

[Cia78] Philippe G. Ciarlet, *The finite element method for elliptic problems*, North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4. MR0520174

[CKS07] Bernardo Cockburn, Guido Kanschat, and Dominik Schötzau, *A note on discontinuous Galerkin divergence-free solutions of the Navier-Stokes equations*, J. Sci. Comput. **31** (2007), no. 1-2, 61–73, DOI 10.1007/s10915-006-9107-7. MR2304270

[Eva98] Lawrence C. Evans, *Partial differential equations*, Graduate Studies in Mathematics, vol. 19, American Mathematical Society, Providence, RI, 1998. MR1625845

[Fix73] George J. Fix, *Eigenvalue approximation by the finite element method*, Advances in Math. **10** (1973), 300–316, DOI 10.1016/0001-8708(73)90113-8. MR0341900

[Gri85] P. Grisvard, *Elliptic problems in nonsmooth domains*, Monographs and Studies in Mathematics, vol. 24, Pitman (Advanced Publishing Program), Boston, MA, 1985. MR775683

[GRS07] Christian Grossmann and Hans-Görg Roos, *Numerical treatment of partial differential equations*, Universitext, Springer, Berlin, 2007. Translated and revised from the 3rd (2005) German edition by Martin Stynes. MR2362757

[GT98] David Gilbarg and Neil S. Trudinger, *Elliptic partial differential equations of second order*, Springer-Verlag, Berlin-New York, 1977. Grundlehren der Mathematischen Wissenschaften, Vol. 224. MR0473443

[GW76] G. H. Golub and J. H. Wilkinson, *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Rev. **18** (1976), no. 4, 578–619, DOI 10.1137/1018113. MR0413456

[Had02] J. Hadamard. Sur les problemes aux derivees partielles et leur signification physique. *Princeton University Bulletin*, 1902.

[KS14] Guido Kanschat and Natasha Sharma, *Divergence-conforming discontinuous Galerkin methods and C^0 interior penalty methods*, SIAM J. Numer. Anal. **52** (2014), no. 4, 1822–1842, DOI 10.1137/120902975. MR3240852

[SF73] Gilbert Strang and George J. Fix, *An analysis of the finite element method*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1973. Prentice-Hall Series in Automatic Computation. MR0443377

[Wil72] J. H. Wilkinson, *Note on matrices with a very ill-conditioned eigenproblem*, Numer. Math. **19** (1972), 176–178, DOI 10.1007/BF01402528. MR0311092

Interdisciplinary Center for Scientific Computing, Heidelberg University, Mathematikon, Klaus-Tschira-Platz 1, 69120 Heidelberg, Germany

E-mail address: kanschat@uni-heidelberg.de