

VARIATIONAL METHODS FOR EIGENVALUE APPROXIMATION

PAT QUILLEN

Abstract. The following is a discussion of the variational characterization of eigenvalues, commonly called the min-max or max-min principle and the Rayleigh-Ritz method. We discuss in some detail the generalized Ritz-Temple inequality which is used to obtain a lower bound on a given eigenvalue, and therefore gives an error estimate on eigenvalue approximations from the Rayleigh-Ritz method.

1. Introduction. The problem we are concerned with here is the eigenvalue problem

$$Au = \lambda u$$

for densely defined self-adjoint operators on a Hilbert space \mathcal{H} . We denote the inner product on \mathcal{H} by $\langle \cdot, \cdot \rangle$, and the associated norm is given by $\|u\| = \sqrt{\langle u, u \rangle}$. The eigenvalues of self-adjoint operators have a variational characterization which has been exploited in order to approximate them. The most notable method based on the variational characterization of the eigenvalues is the Rayleigh-Ritz method, which restricts the operator of interest to a finite dimensional subspace. Eigenvalues of the original operator are then approximated by eigenvalues of the constrained operator.

The Rayleigh-Ritz method alone is somewhat incomplete, as it gives no error estimate for its approximation. Here enters the Temple inequality, put forth first in [10] as an effort to give an error bound on the approximation to eigenvalues given by the Rayleigh-Ritz Method. This inequality does just that by bounding the least eigenvalue below, taking as input an estimate for a lower bound of the second eigenvalue. It's drawback, as cited in [6] is that the Temple formula is not applicable to operators which has densely clustered eigenvalues, as the error is intolerable in this case. In [6], Kato extends the formula to the case of degenerate eigenvalues of operators which are not necessarily semi-bounded below, as we shall typically assume.

Throughout the paper we shall make use of the following form of the spectral theorem for self-adjoint operators as stated in [9]:

THEOREM 1.1. *There is a one-to-one correspondence between self-adjoint operators A and projection-valued measures $\{P_\Omega\}$ on \mathcal{H} , the correspondence being given by*

$$A = \int_{-\infty}^{\infty} \lambda dP_\lambda$$

In proving the max-min principle, we'll also find useful the following related proposition found in [9]:

PROPOSITION 1.2. *Suppose A is self-adjoint and $\lambda \in \mathbb{R}$. Then $\lambda \in \sigma(A)$ if and only if $P_{(\lambda-\varepsilon, \lambda+\varepsilon)} \neq 0$ for all $\varepsilon > 0$.*

The remainder of the paper is organized as follows: In §2 we consider the variational characterization of eigenvalues of a semi-bounded self-adjoint operator, known as the max-min principle. We consider the Rayleigh-Ritz method in §3 and give an example of its use. We treat the Temple inequality and Kato's refinement in §4 of this paper and discuss some of the consequences of the inequality. Finally, concluding remarks are made in §5.

2. Variational Characterization of Eigenvalues. The main problems of the calculus of variations are those which seek to find extrema or stationary points of functionals, just as the calculus may be used to locate extrema of functions. As eigenvectors may be thought of a stationary points of an operator when normalized to lie in the unit sphere, it seems only logical that we may apply methods from the calculus of variations in an attempt to find eigenvectors (or indeed eigenfunctions) of a particular linear operator. In fact, as mentioned in the introduction, eigenvalues of semi-bounded self-adjoint operators have a variational characterization known as the max-min principle ¹.

Throughout this section, we shall make the assumption that the operator A is semi-bounded below. That is, there exists some constant $k \in \mathbb{R}$ such that

$$\langle u, Au \rangle \geq k \langle u, u \rangle$$

for all $u \in \mathcal{D}(A)$. We make this assumption so that the so-called Rayleigh quotient

$$\frac{\langle u, Au \rangle}{\langle u, u \rangle}$$

is itself bounded below, thus insuring the existence of a greatest lower bound.

2.1. The Max-Min Principle. The max-min principle for self-adjoint operators gives a useful characterization of the eigenvalues of the operator. Both the statement and proof of the theorem are adaptations of those found in [8].

THEOREM 2.1 (Max-Min Principle). *Suppose A is a self-adjoint operator which is semi-bounded below. Define*

$$\mu_n(A) := \sup_{\phi_1, \phi_2, \dots, \phi_{n-1}} \left\{ \inf_{\psi \in (V_{n-1}^\perp \cap \mathcal{D}(A)) \setminus \{0\}} \left\{ \frac{\langle \psi, A\psi \rangle}{\langle \psi, \psi \rangle} \right\} \right\}$$

Here $V_{n-1} = \text{span} \{\phi_1, \phi_2, \dots, \phi_{n-1}\}$.

Then, for each fixed n , exactly one of the following holds:

1. There are n eigenvalues (according to multiplicity) below the bottom of the essential spectrum and $\mu_n(A)$ is the n^{th} eigenvalue (again, counting multiplicity).
2. μ_n is the bottom of the essential spectrum and there are at most $n - 1$ eigenvalues below μ_n counting multiplicity.

Proof. By theorem 1.1 there exists a unique family of projection valued measures $\{P_\Omega\}$ such that

$$A = \int_{-\infty}^{\infty} \lambda dP_\lambda$$

We prove that

$$\dim \{\text{ran} \{P_{(-\infty, a)}\}\} < n \quad \text{if } a < \mu_n(A) \quad (*)$$

$$\dim \{\text{ran} \{P_{(-\infty, a)}\}\} \geq n \quad \text{if } a > \mu_n(A) \quad (**)$$

To this end, suppose that $(*)$ is false. That is, suppose that there exists an n -dimensional space $W \subseteq \mathcal{D}(A)$ such that for any $\psi \in W$, $\frac{\langle \psi, A\psi \rangle}{\langle \psi, \psi \rangle} \leq a$, while $a < \mu_n(A)$.

¹This is for operators which are semi-bounded below. Operators which are semi-bounded above have the min-max principle.

That we may assume $W \subseteq \mathcal{D}(A)$ follows from the assumption that A is semi-bounded below. Now given any collection $\{\phi_1, \phi_2, \dots, \phi_{n-1}\}$ (whose span is the space V_{n-1}), there exists $\psi \in W \cap V_{n-1}^\perp$ such that $\psi \neq 0$, since $\dim \{V_{n-1}\} \leq n-1 < n = \dim \{W\}$. As $\psi \in W$, we have that

$$\inf_{\psi \in (V_{n-1}^\perp \cap \mathcal{D}(A)) \setminus \{0\}} \left\{ \frac{\langle \psi, A\psi \rangle}{\langle \psi, \psi \rangle} \right\} \leq a$$

As our space V_{n-1} was arbitrary, it readily follows that $\mu_n(A) \leq a$, contradicting the assumption that $a < \mu_n$.

Now, suppose that (**) is false. That is, suppose that $\dim \{\text{ran } \{P_{(-\infty, a)}\}\} < n$, while $a > \mu_n(A)$. In particular, suppose that there exists a collection $\{\hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_{n-1}\}$ such that $\text{ran } \{P_{(-\infty, a)}\} = \text{span } \{\hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_{n-1}\} = \hat{V}_{n-1}$. Therefore, any element $\psi \in \hat{V}_{n-1}^\perp \cap \mathcal{D}(A)$ is in fact an element of $\text{ran } \{P_{[a, \infty)}\}$ (this is readily seen by appealing to theorem 1.1). Therefore, we get that

$$\frac{\langle \psi, A\psi \rangle}{\langle \psi, \psi \rangle} \geq a \quad \text{for all } \psi \in \hat{V}_{n-1}^\perp \cap \mathcal{D}(A).$$

Thus $\mu_n(A) \geq a$ contradicting the hypothesis that $a > \mu_n(A)$.

Together, (*) and the fact that A is semi-bounded below imply that $\mu_n(A)$ is finite for fixed n . We now examine two possible cases. Either $\text{ran } \{P_{(-\infty, \mu_n(A)+\varepsilon)}\}$ is an infinite dimensional subspace of \mathcal{H} for all $\varepsilon > 0$ or there exists $\hat{\varepsilon} > 0$ such that the range of the associated projector is a finite dimensional space. We consider the second case first.

Case 1. $\dim \{\text{ran } \{P_{(-\infty, \mu_n(A)+\hat{\varepsilon})}\}\} < \infty$ for some $\hat{\varepsilon} > 0$

By (*) and (**) together, we have that $\dim \{\text{ran } \{P_{(\mu_n(A)-\varepsilon, \mu_n(A)+\varepsilon)}\}\} \geq 1$ for all $\varepsilon > 0$, and therefore by proposition (1.2), we have that $\mu_n(A) \in \sigma(A)$. Also, by our supposition that $\dim \{\text{ran } \{P_{(-\infty, \mu_n(A)+\hat{\varepsilon})}\}\} < \infty$ for some $\hat{\varepsilon} > 0$, it follows that $\mu_n(A)$ is an eigenvalue of A . As $\mu_n(A)$ is an eigenvalue of A , there exists $r > 0$ such that $\sigma(A) \cap B(\mu_n(A), r) = \{\mu_n(A)\}$, and so we see that, by (**), $\dim \{\text{ran } \{P_{(-\infty, \mu_n(A))}\}\} = \dim \{\text{ran } \{P_{(-\infty, \mu_n(A)+r)}\}\} \geq n$. Therefore, there are at least n eigenvalues, counting multiplicity, which are less than or equal to $\mu_n(A)$. If in fact, E_n , the n^{th} eigenvalue of A , is strictly less than $\mu_n(A)$, we get that $\dim \{\text{ran } \{P_{(-\infty, E_n)}\}\} = n$, as E_n is an eigenvalue of A . Simultaneously, by (*) we have that $\dim \{\text{ran } \{P_{(-\infty, E_n)}\}\} < n$ as $E_n < \mu_n(A)$. This contradiction shows that we are in situation 1 of the theorem.

Case 2. $\dim \{\text{ran } \{P_{(-\infty, \mu_n(A)+\varepsilon)}\}\} = \infty$ for all $\varepsilon > 0$

From (*), we have that $\dim \{\text{ran } \{P_{(-\infty, \mu_n(A)-\varepsilon)}\}\} \leq n-1$, and therefore it follows that $\dim \{\text{ran } \{P_{(\mu_n(A)-\varepsilon, \mu_n(A)+\varepsilon)}\}\} = \infty$ for all $\varepsilon > 0$. Thus, by definition, $\mu_n(A) \in \sigma_{\text{ess}}(A)$. Also, if $a < \mu_n(A)$ and $0 < \varepsilon < \mu_n(A) - a$, then from (*) it follows that $\dim \{\text{ran } \{P_{(a-\varepsilon, a+\varepsilon)}\}\} < n < \infty$. Therefore, by letting $a \rightarrow \mu_n(A)$, we get that $\mu_n(A) = \inf \{\lambda : \lambda \in \sigma_{\text{ess}}(A)\}$. Furthermore, it follows that $\mu_{n+1}(A) \geq \mu_n(A)$, merely by the definition of $\mu_n(A)$. Also, we have equality here, enforced by the condition (*). Similarly, if we suppose that there are n or more eigenvalues less than $\mu_n(A)$, we arrive at a contradiction, again from condition (*). Clearly, we are in situation 2 of the theorem, and the proof is complete. \square

Now if A and B are both self-adjoint operators which are semi-bounded below with the property that $\mathcal{D}(B) \subseteq \mathcal{D}(A)$ and

$$\langle \phi, A\phi \rangle \leq \langle \phi, B\phi \rangle \quad \text{for all } \phi \in \mathcal{D}(B)$$

then we say that $A \leq B$. The following is an immediate corollary of the max-min principle.

COROLLARY 2.2. *If A and B are two operators such that $A \leq B$, then*

$$\mu_n(A) \leq \mu_n(B)$$

Proof. By the max-min principle,

$$\begin{aligned} \mu_n(B) &= \sup_{\phi_1, \phi_2, \dots, \phi_{n-1}} \left\{ \inf_{\psi \in (V_{n-1}^\perp \cap \mathcal{D}(B)) \setminus \{0\}} \left\{ \frac{\langle \psi, B\psi \rangle}{\langle \psi, \psi \rangle} \right\} \right\} \\ &\geq \sup_{\phi_1, \phi_2, \dots, \phi_{n-1}} \left\{ \inf_{\psi \in (V_{n-1}^\perp \cap \mathcal{D}(B)) \setminus \{0\}} \left\{ \frac{\langle \psi, A\psi \rangle}{\langle \psi, \psi \rangle} \right\} \right\} \\ &\geq \sup_{\phi_1, \phi_2, \dots, \phi_{n-1}} \left\{ \inf_{\psi \in (V_{n-1}^\perp \cap \mathcal{D}(A)) \setminus \{0\}} \left\{ \frac{\langle \psi, A\psi \rangle}{\langle \psi, \psi \rangle} \right\} \right\} = \mu_n(A) \end{aligned}$$

since $A \leq B$. \square

This particular corollary will prove useful in establishing a lower bound on the second eigenvalue, which is required for Temple's inequality.

3. Rayleigh-Ritz Method. The max-min principle is the foundation of the Rayleigh-Ritz method for eigenvalue approximation, and in fact, the Rayleigh-Ritz method provides upper bounds on the eigenvalues of the operator under consideration, by restricting the operator to a finite dimensional space. Physically, the more constraints that are placed on a vibrating system causes the fundamental frequency and all subsequent overtones to be raised, or at least not lowered (see [3] [7]). That is, the more constraints we place upon the operator, such as restricting it to certain subspaces of its space of definition, requires more energy, and thus raises the eigenvalues. In fact, as seen in [3] and [5], for the finite-dimensional case, the eigenvalues of any $r \times r$ principal submatrix of an $n \times n$ hermitian matrix A are interlaced with the eigenvalues of the matrix itself. That is, for all $k : 1 \leq k \leq r$, we have that

$$\lambda_k(A) \leq \lambda_k(A_r) \leq \lambda_{k+n-r}(A)$$

In particular, for $r = n - 1$, we have

$$\mu_1 \leq \lambda_1 \leq \mu_2 \leq \lambda_2 \leq \dots \leq \mu_k \leq \lambda_k \leq \mu_{k+1} \leq \dots \leq \lambda_{n-1} \leq \mu_n$$

where $\{\lambda_i\}_{i=1}^{n-1}$ are the eigenvalues of A_{n-1} and $\{\mu_i\}_{i=1}^n$ are the eigenvalues of A . This is the essence of the Rayleigh-Ritz method.

THEOREM 3.1 (Rayleigh-Ritz Method). *Suppose A is a self-adjoint operator semi-bounded below. Let V be an n -dimensional subspace of $\mathcal{D}(A)$ and let P be the orthogonal projection onto V . Let $A_V = PAP$ with eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ such that $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. Then*

$$\mu_m(A) \leq \lambda_m, \quad m = 1, \dots, n$$

In particular, if A has eigenvalues E_1, E_2, \dots, E_k , counting multiplicity, at the bottom of its spectrum, then

$$E_m \leq \lambda_m, \quad m = 1, \dots, \min\{k, n\}$$

Proof. Begin by noting that since P is an orthogonal projection onto V , we have that $\langle \psi, P\phi_i \rangle = \langle P\psi, \phi_i \rangle = \langle \psi, \phi_i \rangle$ for all $\psi \in V$. Thus

$$(PV_{m-1})^\perp \cap V = (\text{span}\{P\phi_1, P\phi_2, \dots, P\phi_{m-1}\})^\perp \cap V = V_{m-1}^\perp \cap V$$

Also, as $V \subseteq \mathcal{H}$, it follows immediately that $\inf_{\psi \in V} \{f(\psi)\} \geq \inf_{\psi \in \mathcal{H}} \{f(\psi)\}$. And so, by the max-min principle, the eigenvalues of A_V are given by

$$\begin{aligned} \lambda_m &= \sup_{\phi_1, \phi_2, \dots, \phi_{m-1} \in V} \left\{ \inf_{\psi \in (V_{m-1}^\perp \cap V) \setminus \{0\}} \left\{ \frac{\langle \psi, A_V \psi \rangle}{\langle \psi, \psi \rangle} \right\} \right\} \\ &= \sup_{\phi_1, \phi_2, \dots, \phi_{m-1} \in \mathcal{H}} \left\{ \inf_{\psi \in ((PV_{m-1})^\perp \cap V) \setminus \{0\}} \left\{ \frac{\langle \psi, A \psi \rangle}{\langle \psi, \psi \rangle} \right\} \right\} \\ &= \sup_{\phi_1, \phi_2, \dots, \phi_{m-1} \in \mathcal{H}} \left\{ \inf_{\psi \in (V_{m-1}^\perp \cap V) \setminus \{0\}} \left\{ \frac{\langle \psi, A \psi \rangle}{\langle \psi, \psi \rangle} \right\} \right\} \\ &\geq \sup_{\phi_1, \phi_2, \dots, \phi_{m-1} \in \mathcal{H}} \left\{ \inf_{\psi \in (V_{m-1}^\perp \cap \mathcal{D}(A)) \setminus \{0\}} \left\{ \frac{\langle \psi, A \psi \rangle}{\langle \psi, \psi \rangle} \right\} \right\} \\ &= \mu_m(A) \end{aligned}$$

and the theorem is proved. \square

The Rayleigh-Ritz method is frequently applied to approximate the eigenvalues of large sparse linear systems. In the case of hermitian matrices, the Lanczos method is applied to build an $n \times k$ matrix Q_k with orthonormal columns so that

$$AQ_k = Q_k T_k + \beta q_{k+1} e_k^T$$

where T_k is a tridiagonal matrix. The eigenvalues of T_k , called Ritz values for A , are then used to approximate the eigenvalues of A . See [4] for more details concerning such approximations and their accuracy. The example that follows, however, shows a more classical application of the Rayleigh-Ritz method.

3.1. Example: One-dimensional Harmonic Oscillator. The Schrödinger operator for a one-dimensional harmonic oscillator, given by

$$H = -\frac{d^2}{dx^2} + \omega^2 x^2$$

is a nonnegative self-adjoint operator, and therefore, we may apply the Rayleigh-Ritz method to get a bound for its ground state energy $\mu_1(H)$. Noting that $\phi_\alpha = e^{-\alpha x^2} \in \mathcal{D}(H)$ for any $\alpha > 0$ we have that

$$\mu_1(H) \leq \min_{\alpha > 0} \left\{ \frac{\langle \phi_\alpha, H \phi_\alpha \rangle}{\langle \phi_\alpha, \phi_\alpha \rangle} \right\}$$

Now then

$$H\phi_\alpha = 2\alpha e^{-\alpha x^2} + (\omega^2 x - 4\alpha^2)x^2 e^{-\alpha x^2}$$

and thus

$$\begin{aligned}\langle \phi_\alpha, H\phi_\alpha \rangle &= \int_{\mathbb{R}} (e^{-\alpha x^2})(2\alpha e^{-\alpha x^2} + (\omega^2 x - 4\alpha^2)x^2 e^{-\alpha x^2}) dx \\ &= \left(2\alpha + \frac{\omega^2 - 4\alpha^2}{4\alpha}\right) \|\phi_\alpha\|^2\end{aligned}$$

Therefore

$$\frac{\langle \phi_\alpha, H\phi_\alpha \rangle}{\langle \phi_\alpha, \phi_\alpha \rangle} = \frac{4\alpha^2 + \omega^2}{4\alpha} = \alpha + \frac{\omega^2}{4\alpha}$$

Minimizing with respect to α finds that an absolute minimum is achieved at $\alpha = \frac{\omega}{2}$, and therefore

$$\mu_1(H) \leq \frac{\omega}{2} + \frac{\omega^2}{2\omega} = \omega$$

In fact, $\phi = e^{-\frac{\omega}{2}x^2} \in \mathcal{D}(H)$ is an eigenfunction with the eigenvalue ω . That ω is in fact the ground state energy is easily verified. Regardless of this, the Rayleigh-Ritz method guarantees that $\mu_1(H) \leq \omega$.

4. Temple's Inequality: A Lower Bound on the Ground State Energy.

In an attempt to get an error bound for approximations of eigenvalues by the Rayleigh-Ritz method, Temple first put forth the following inequality in [10]:

THEOREM 4.1. *Suppose that $\{\mu_n\}$ is a strictly decreasing sequence bounded below by λ_1 , the ground state energy of the operator under consideration. Then*

$$\lambda_1 > \mu_n - \frac{\mu_{n-1} - \mu_n}{\frac{\lambda_2}{\mu_n} - 1}$$

or as stated in [8]

THEOREM 4.2. *Suppose that $\check{\mu} : \lambda_1 < \check{\mu} < \lambda_2$, and $\phi \in \mathcal{D}(A)$ such that $\|\phi\| = 1$. Then*

$$\lambda_1 > \langle \phi, A\phi \rangle - \frac{\|A\phi\|^2 - \langle \phi, A\phi \rangle^2}{\check{\mu} - \langle \phi, A\phi \rangle}$$

We omit the proof of this theorem, and proceed to Kato's generalization found in [6]. As we shall see, Kato strengthened the inequality considerably by removing all restrictions on the operator save self-adjointness. Namely, the generalized Ritz-Temple formula does not require that the operator be semi-bounded below. Instead, to use the formula we must provide an open interval (a, b) which contains a single isolated eigenvalue and no other piece of the spectrum.

4.1. Background. In order to prove the generalized Ritz-Temple inequality, we require following proposition:

PROPOSITION 4.3. *Suppose that A is self-adjoint and $(a, b) \subseteq \mathbb{R}$ such that $(a, b) \cap \sigma(A) = \emptyset$. Suppose further that $\phi \in \mathcal{D}(A)$ with $\|\phi\| = 1$. Then*

$$\left\| \left(A - \frac{a+b}{2} \right) \phi \right\|^2 \geq \left(\frac{a-b}{2} \right)^2$$

Proof. We note the following equivalence:

$$\begin{aligned} \left\| \left(A - \frac{a+b}{2} \right) \phi \right\|^2 - \left(\frac{a-b}{2} \right)^2 &= \langle A\phi, A\phi \rangle - (a+b) \langle \phi, A\phi \rangle - ab \langle \phi, \phi \rangle \\ &= \langle (A-a)\phi, (A-b)\phi \rangle \end{aligned}$$

Since A is self-adjoint, it readily follows that each of $(A-a)$ and $(A-b)$ are themselves self-adjoint, and moreover, as a consequence of theorem 1.1 we have that

$$\langle (A-a)\phi, (A-b)\phi \rangle = \int_{-\infty}^{\infty} (\lambda-a)(\lambda-b) d \langle \phi, P_\lambda \phi \rangle$$

As $(a, b) \cap \sigma(A) = \emptyset$, it follows that $\langle \phi, P_{(a,b)} \phi \rangle = 0$ and so

$$\langle (A-a)\phi, (A-b)\phi \rangle = \int_{\mathbb{R} \setminus (a,b)} (\lambda-a)(\lambda-b) d \langle \phi, P_\lambda \phi \rangle$$

Here the integrand is always nonnegative and since P_Ω is an orthogonal projection for any Borel set Ω it readily follows that the measure $\langle \phi, P_\Omega \phi \rangle$ is a positive Borel measure for fixed ϕ . Therefore, $\langle (A-a)\phi, (A-b)\phi \rangle \geq 0$. \square

For a given $\phi \in \mathcal{D}(A)$ with $\|\phi\| = 1$ we define

$$\gamma := \langle \phi, A\phi \rangle \quad \text{and} \quad \eta := \langle A\phi, A\phi \rangle - \langle \phi, A\phi \rangle^2 = \|A\phi\|^2 - \gamma^2$$

Note that

$$\begin{aligned} \|(A-\gamma)\phi\|^2 &= \langle (A-\gamma)\phi, (A-\gamma)\phi \rangle \\ &= \langle A\phi, A\phi \rangle - 2\gamma \langle \phi, A\phi \rangle + \gamma^2 \langle \phi, \phi \rangle \\ &= \|A\phi\|^2 - \gamma^2 = \eta \end{aligned}$$

So η is the residual norm squared. That is, η is effectively a measure of how far the pair (γ, ϕ) is from being an eigenpair for the operator A . The significance of this will be discussed more fully later.

Now then, with the above definitions, we have that

$$\begin{aligned} \langle (A-a)\phi, (A-b)\phi \rangle &= \langle A\phi, A\phi \rangle - (a+b) \langle \phi, A\phi \rangle + ab \\ &= \eta + \gamma^2 - (a+b)\gamma + ab \end{aligned}$$

If it is the case that $\langle (A-a)\phi, (A-b)\phi \rangle \geq 0$ then

$$\begin{aligned} \eta &\geq -\gamma^2 + (a+b)\gamma - ab \\ &= (\gamma-a)(-\gamma) + (\gamma-a)b \\ &= (\gamma-a)(b-\gamma) \end{aligned}$$

Therefore, by proposition 4.3, it follows that if $(a, b) \cap \sigma(A) = \emptyset$, then $\eta \geq (\gamma-a)(b-\gamma)$. More importantly, the contrapositive of this statement holds, and we shall make use of this fact in proving the generalized Ritz-Temple formula.

4.2. The Generalized Ritz-Temple Formula. What follows is the generalized Ritz-Temple formula proved by Kato in [6]. We present a slightly different version of the theorem. We make no assumptions on the degeneracy of the eigenvalue found in the bracketing interval. This differs from the theorem of Kato which required that the eigenvalue which is approximated by the formula be simple so that an estimate concerning the associated eigenvector may be made. We include no such estimate here and therefore do not require that the eigenvalue lying in the interval (a, b) be simple. The inequality does not give much in the way of a good bound when considering a collection of densely clustered eigenvalues, and Kato extends the inequality to deal with this situation. We will not discuss this result here and point the interested reader to [6].

THEOREM 4.4 (Generalized Ritz-Temple Inequality). *Suppose that A is a self-adjoint operator and $(a, b) \subseteq \mathbb{R}$ such that $(a, b) \cap \sigma(A) = \{\lambda\}$, a single point. Suppose further that $\phi \in \mathcal{D}(A)$ with $\|\phi\| = 1$. With γ and η as above, if*

$$\eta < (\gamma - a)(b - \gamma) \tag{†}$$

then

$$\gamma - \frac{\eta}{b - \gamma} \leq \lambda \leq \gamma + \frac{\eta}{\gamma - a}$$

Proof. As $\eta = \|(A - \gamma)\phi\|^2$, it follows immediately that $\eta \geq 0$. Now, by (†) and the definitions of η and γ , we have

$$\begin{aligned} \langle A\phi, A\phi \rangle - \langle \phi, A\phi \rangle^2 &< (\langle \phi, A\phi \rangle - a)(b - \langle \phi, A\phi \rangle) \\ &= \langle \phi, A\phi \rangle (b - a) - ab - \langle \phi, A\phi \rangle^2 \end{aligned}$$

or equivalently

$$ab < ab + \langle A\phi, A\phi \rangle < \langle \phi, A\phi \rangle (b - a)$$

which reduces to

$$\frac{ab}{b - a} < \langle \phi, A\phi \rangle = \gamma$$

That is, $\gamma > a$ if $\frac{ab}{b-a} \geq a$. Showing that $\frac{ab}{b-a} \geq a$ is trivial, and so, $\gamma > a$. As $\eta \geq 0$, (†) implies that $\gamma < b$ as well.

Now, put

$$a' := \gamma - \frac{\eta}{b - \gamma} \quad \text{and} \quad b' := \gamma + \frac{\eta}{\gamma - a}$$

By (†) it follows that

$$\begin{aligned} a' &= \gamma - \frac{\eta}{b - \gamma} > \gamma - (\gamma - a) = a \\ &\quad \text{and} \\ b' &= \gamma + \frac{\eta}{\gamma - a} < \gamma + (b - \gamma) = b \end{aligned}$$

Also, since $\eta \geq 0$ and $a < \gamma < b$, we have that

$$a < a' \leq \gamma \leq b' < b$$

Now then, for any $a'' : a < a'' < a'$ we have that $(a'' - \gamma) < \frac{-\eta}{b-\gamma}$ and so, in fact,

$$\eta < (\gamma - a'')(b - \gamma)$$

Therefore, by proposition 4.3, it follows that $(a'', b) \cap \sigma(A) \neq \emptyset$.

Repeating the argument above, we see that we may select a number b'' such that $b' < b'' < b$ and $(a, b'') \cap \sigma(A) \neq \emptyset$. Since we know that $(a, b) \cap \sigma(A) = \{\lambda\}$, a single isolated eigenvalue, we see that $(a'', b'') \cap \sigma(A) \neq \emptyset$. As a'' and b'' are arbitrary, then, we have that $[a', b'] \cap \sigma(A) \neq \emptyset$ and the theorem is proved. \square

As a consequence of theorem 4.4 then,

$$|\lambda - \gamma| \leq \frac{\|r\|^2}{g} \quad (\ddagger)$$

where $r = (A - \gamma)\phi$ is the residual and $g = \min \{\gamma - a, b - \gamma\}$. We see that (\ddagger) implies that the error in γ depends on the residual norm *squared*, as well as the gap g . This has two consequences. First, if the gap is suitably large, then the Rayleigh quotient will converge rapidly² to the eigenvalue in the finite dimensional case. Secondly, if the gap is small, that is, the eigenvalues of interest are clustered closely together, then regardless of the residual norm, the error can still be large.

4.3. Example: Applying the generalized Ritz-Temple Formula. This example appears in [6]. Consider the operator

$$\tilde{H} = -\frac{d^2}{dx^2} + x^4$$

It is readily seen that this operator is self-adjoint and for any $\alpha > 0$ we have that

$$\phi_\alpha = \left(\frac{\alpha}{\pi}\right)^{\frac{1}{4}} e^{-\frac{1}{2}\alpha x^2} \in \mathcal{D}(\tilde{H})$$

and $\|\phi_\alpha\| = 1$.

We need to calculate $\gamma = \langle \phi_\alpha, \tilde{H}\phi_\alpha \rangle$ and $\eta = \langle \tilde{H}\phi_\alpha, \tilde{H}\phi_\alpha \rangle - \gamma^2$. Note that

$$\tilde{H}\phi_\alpha = \left(\frac{\alpha}{\pi}\right)^{\frac{1}{4}} (\alpha - \alpha^2 x^2 + x^4) e^{-\frac{1}{2}\alpha x^2}$$

We therefore have

$$\gamma = \frac{1}{2}\alpha + \frac{3}{4}\alpha^{-2} \quad \text{and} \quad \eta = \frac{1}{2}\alpha^2 - 3\alpha^{-1} + 6\alpha^{-4}$$

To apply the generalized Ritz-Temple inequality, we must bracket the first eigenvalue. We note that $(x^2 - \frac{1}{2}\omega^2)^2 \geq 0$ for all $x \in \mathbb{R}$ and any choice of $\omega > 0$. In particular,

$$\tilde{H} \geq H - \frac{1}{4}\omega^4$$

²Convergence is locally cubic in fact. That is, once the condition (\ddagger) is met, approximations to the eigenvalue λ gain roughly three correct digits at each iteration of Rayleigh quotient iteration (RQI). This cubic convergence is a direct consequence of the error estimate depending on the residual norm squared. For a more complete discussion of RQI and this property in particular, consult [4].

where H is the Schrödinger operator for the one-dimensional harmonic oscillator.

By corollary 2.2, then

$$\mu_2(\tilde{H}) \geq \mu_2(H) - \frac{1}{4}\omega^4 = 3\omega - \frac{1}{4}\omega^4$$

Maximizing the right hand side with respect to ω finds that

$$\mu_2(\tilde{H}) \geq \frac{3}{4}3^{\frac{4}{3}}$$

gives the tightest lower bound for the second eigenvalue. Thus, we take $b = \frac{3}{4}3^{\frac{4}{3}}$. As \tilde{H} is a nonnegative operator, we may as well take $a = -\infty$.

Then by theorem 4.4

$$\frac{1}{2}\alpha + \frac{3}{4}\alpha^{-2} - \frac{\frac{1}{2}\alpha^2 - 3\alpha^{-1} + 6\alpha^{-4}}{\frac{3}{4}3^{\frac{4}{3}} - (\frac{1}{2}\alpha + \frac{3}{4}\alpha^{-2})} \leq \mu_1(\tilde{H}) \leq \frac{1}{2}\alpha + \frac{3}{4}\alpha^{-2}$$

An approximate minimization of the right hand side and an approximate maximization of the left hand side give

$$0.9459 \leq \mu_1(\tilde{H}) \leq 1.082$$

5. Conclusion. We have considered variational methods for eigenvalue approximations. Most importantly, we have discussed the accuracy of the approximations taken from these variational characterizations. We have seen that the generalized Ritz-Temple formula allows one to use the Rayleigh quotient to approximate not just simple eigenvalues, but degenerate eigenvalues to a high degree of accuracy provided that the condition (\dagger) holds and the eigenvalue is sufficiently isolated from the rest of the spectrum. Allowing the degeneracy of the eigenvalue approximated, however, has the undesirable, but natural side effect of obtaining a poor approximate associated eigenvector. Some preliminary numerical experiments seem to indicate that in such a situation the vector ϕ which masquerades as an approximate eigenvector in fact approximately lies in the eigenspace associated with the eigenvalue approximated by γ . We expect that there is a good amount of literature concerning the traditional Rayleigh quotient iteration supporting this conjecture, however, sadly, we are unaware of it at this time. As a future project, we intend to investigate this phenomenon and also look into methods such as the Grassmann-Rayleigh quotient iteration proposed in [1] which strive to compute invariant subspaces of the operator concerned.

REFERENCES

- [1] P.-A. ABSIL, R. MAHONY, R. SEPULCHRE, AND P. V. DOOREN, *A Grassmann-Rayleigh Quotient Iteration for Computing Invariant Subspaces*, SIAM Review, 44 (2002), pp. 57–73.
- [2] R. COURANT, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc., 49 (1943), pp. 1–23.
- [3] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics, Volume 1*, Wiley Interscience, New York, 1953.
- [4] J. W. DEMMEL, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
- [5] R. A. JOHNSON AND C. R. HORN, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.
- [6] T. KATO, *On the upper and lower bounds of eigenvalues*, J. Phys. Soc. Japan, 4 (1949), pp. 334–339.
- [7] RAYLEIGH, *Theory of Sound*, Dover, New York, 2 ed., 1945.

- [8] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics IV: Analysis of Operators*, Academic Press, 1978.
- [9] ———, *Methods of Modern Mathematical Physics I: Functional Analysis*, Academic Press, 1980.
- [10] G. TEMPLE, *The theory of rayleigh's principle as applied to continuous systems*, Proc. Roy. Soc. London, 119A (1928), pp. 276–293.
- [11] H. F. WEINBERGER, *Variational Methods for Eigenvalue Approximation*, SIAM, Philadelphia, 1974.