BIT manuscript No.
(will be inserted by the editor)

Optimal quotients for solving large eigenvalue problems

Marko Huhtanen · **Vesa Kotila**

the date of receipt and acceptance should be inserted later

Abstract Quotients for eigenvalue problems (generalized or not) are considered. To have a quotient optimally approximating an eigenvalue, conditions are formulated to maximize the one-dimensional projection of the eigenvalue problem. Respective optimal quotient iterations are derived under the assumption that applying the inverse is affordable. Inexact methods are also considered if applying the inverse is not affordable. Then, to approximate an eigenvector, optimality conditions are formulated to minimize linear independency over a subspace. Equivalence transformations are performed for preconditioning iterations and steering the convergence. These ideas extend to subspaces in a natural way. For the standard eigenvalue problem, a new Arnoldi method arises as an alternative to the classical Arnoldi method.

Keywords Generalized eigenvalue problem · Optimal quotient · Rayleigh quotient · Field of values · Iterative methods · Cubic convergence · Finite section method · Arnoldi method

Mathematics Subject Classification (2000) 65F15, 15A22, 15A60

1 Introduction

Let the matrices $M,N \in \mathbb{C}^{n\times n}$ be large and possibly sparse. Without assuming any additional structure, a common approach to numerically compute a few eigenpairs of

M. Huhtanen
Faculty of Information Technology and Electrical Engineering, University of Oulu,
90570 Oulu 57, Finland,
Tel.: +358 294 482663
E-mail: Marko.Huhtanen@aalto.fi

V. Kotila
Faculty of Information Technology and Electrical Engineering, University of Oulu,
90570 Oulu 57, 6 Finland,
Tel.: +358 294 482668
E-mail: Vesa.Kotila@oulu.fi

the eigenvalue problem

$$Mx = \lambda Nx \tag{1.1}$$

consists of a computation of an approximate unit eigenvector $q \in \mathbb{C}^n$, or possibly a few.[1] If q is not an eigenvector, then Mq and Nq remain linearly independent. To determine an approximate eigenvalue, called a quotient, project the problem as

$$z^*Mq = \lambda z^*Nq \tag{1.2}$$

by appropriately constructing a unit vector $z \in \mathbb{C}^n$. For a classical option resulting from a study of quadratic forms, choosing $z = q$ yields the Rayleigh quotient; see, e.g., [26], [30, Chapter 4] and [9, Chapter 8]. See in particular [25] and [20, Section 8] for its history and concise descriptions of its origins in physics. (For an overview and a wealth of references on the eigenvalue problem, see [10], [31], [18] and [4].) For large scale computational aspects in finite element modelling in mechanics, see [3]. This paper is concerned with devising an optimal projection (1.2) and resulting iterative methods to solve the eigenvalue problem (1.1).

To optimally construct z, consider maximizing the projection onto the elements of the Grassmannian $\mathrm{Gr}_1(\mathbb{C}^n)$ corresponding to the vectors Nq and Mq.[2] This option leads to choosing z yielding the quotient

$$\frac{q^*N^*Mq}{|q^*N^*Mq|}\frac{\|Mq\|}{\|Nq\|} \tag{1.3}$$

whenever $q^*N^*Mq \neq 0$. Thereafter, in terms of z and this quotient, the approximate eigenvector q can be updated. When repeated, this two-step construction gives rise to a back-and-forth optimal quotient iteration for solving the eigenvalue problem (1.1); see Algorithm 1 in Section 3. It is shown that cubic convergence is attained under the additional assumptions that UMY^{-1} and UNY^{-1} be diagonal for a unitary U and an invertible matrix Y. This provides an intriguing class of generalized eigenvalue problems. In particular, for the standard Hermitian eigenvalue problem, which obviously satisfies these assumptions, this iteration can be argued (and shown by examples) to yield more accurate approximations than the classical Rayleigh quotient iteration.[3] For the Rayleigh quotient iteration, see [25] and [26, Chapter 4].

Equivalence transformations are used for preconditioning to steer the convergence of this basic algorithm to a desired part of the spectrum. When (exact) inversion is dynamically performed from the left, the approximate eigenvector elegantly takes the form of

$$\prod_{j=1}^{k}(M - l_jN)^{-1}(M + l_jN)q,$$

where l_j are the respective optimal quotients; see Algorithm 2. In particular, if N is invertible, then $(M - l_jN)^{-1}(M + l_jN) = (N^{-1}M - l_jI)^{-1}(N^{-1}M + l_jI)$, so that the

[1] For the standard eigenvalue problem, there is the spectacularly simple option to execute the power method to have an approximate eigenvector q corresponding to a dominant eigenvalue.

[2] $\mathrm{Gr}_1(\mathbb{C}^n)$ is the set of one dimensional subspaces of $\mathbb{C}^n$.

[3] This is striking since, e.g., in [4] it is (unfoundedly) claimed that the Rayleigh quotient iteration is "best".

eigenvector approximation evolves in terms of consecutive applications of Cayley transformations. For the use of Cayley transformations and rational methods more generally, see [19] and [28,29] and references therein.

The convergence can further be affected by introducing an optimality criterion for producing an approximate eigenvector. It consists of maximizing the modulus of

$$\frac{v^*N^*Mv}{\|Nv\|\|Mv\|} \tag{1.4}$$

for unit vectors v restricted to the subspace generated through (possibly inaccurate) inversion. Since this is a natural task for the eigenvalue problem in general, and for inexact methods in particular, efficient solving is of particular interest. When repeated to expand the subspace, we have optimal methods for numerically solving (1.1); see Algorithms 4 and 5. This ensures that the convergence becomes monotonic. Altogether, as (1.3) and (1.4) show, the matrix N^*M plays a seemingly important role in the construction of optimal approximations for the generalized eigenvalue problem. Of course, accurate application of the inverse is often unrealistic. Involving inversion and, most likely, iterative methods with preconditioning, exact updating described can become overly expensive. Inexact methods are therefore considered.

Besides single vector projections (1.2) and iterations, the notion of quotient has a natural extension to involve subspaces; see [25, Chapter 11.3] how this extension is done in the case of the Rayleigh quotient for the Hermitian eigenvalue problem. (For the finite section, i.e., Galerkin method for eigenvalue approximations in full generality, see [1].) In particular, for the standard eigenvalue problem this construction implies that the Galerkin method cannot be regarded as optimal.[4] This means, for example, that the classical Arnoldi method gets replaced with an optimal Arnoldi method.

The paper is organized as follows. In Section 2 the notion of optimal quotient is derived. For an eigenvalue inclusion set, their union yields the so-called field of optimal quotients for the eigenvalue problem. Basic properties of the field of optimal quotients are demonstrated. In Section 3, several optimal quotient iterations for solving the eigenvalue problem are introduced. Optimality conditions are posed for overcoming the possible lack of accuracy in performing the inversion with inexact methods. In Section 4 the idea of optimal projection is extended to involve subspaces of dimension larger than one. Section 5 is devoted to numerical experiments.

2 Optimal quotients for approximating eigenvalues

Consider the eigenvalue problem (1.1) under the assumption that the corresponding matrix subspace

$$\mathscr{V} = \operatorname{span}\{M,N\} \tag{2.1}$$

is nonsingular, i.e., contains invertible elements. The choice of a basis of $\mathscr{V}$ is a very important issue in the numerical solution of (1.1). This is due to the equivalence

[4] The difference resembles that of the GMRES and FOM methods for solving linear systems.

of solving the eigenvalue problem when M and N are replaced with any linearly independent linear combinations

$$A = aM + bN \text{ and } B = cM + dN \tag{2.2}$$

of M and N with $a,b,c,d \in \mathbb{C}$. (For example, regarding the choice in the standard Hermitian eigenvalue problem, see [26, Chapter 15.3].) Appropriate linear combinations allow, e.g., making any of the eigenvalues on the outer boundary dominating, i.e., largest in modulus.[5]

Proposition 2.1 *Let $M,N \in \mathbb{C}^{n\times n}$ and suppose* (2.2) *with* $\det\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \neq 0$. *Then the matrix $\alpha A + \beta B$ is singular if and only if $\delta M + \gamma N$ is singular, where* $\left[\begin{smallmatrix} \delta \\ \gamma \end{smallmatrix}\right] = \left[\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right] \left[\begin{smallmatrix} \alpha \\ \beta \end{smallmatrix}\right]$.

In this section we assume the basis has been fixed such that we are concerned with the formulation (1.1).

2.1 Construction of an optimal quotient

With two unit vectors available, consider the task of approximating eigenvalues.

Definition 2.1 The complex number λ satisfying (1.2) is said to be the quotient of the eigenvalue problem (1.1) corresponding to the unit vectors z and q.

Both z and q should be carefully constructed. Resulting from a study of quadratic forms, a classical alternative consists of choosing $z = q$. Then it is common to assume M to be Hermitian and N Hermitian positive definite. This choice leads to the Rayleigh quotient and the respective notion of field of values; see [13, Chapter 22], [15, Chapter 1], [26, Chapter 15] and [12]. For more recent studies, see [27] and [6] and references therein. For references on developments in Banach spaces, see [21].

For the generalized eigenvalue problem the choice $z = q$ is not arguable in general. A minimum criterion is that if q is an eigenvector, then the quotient should always yield the corresponding eigenvalue. The choice $z = q$ does not satisfy this in general.[6] To have this property in terms of an optimal construction, let $w_1 = \frac{Mq}{\|Mq\|}$ and $w_2 = \frac{Nq}{\|Nq\|}$ be nonzero. If q is not an eigenvector, then w_1 and w_2 are linearly independent. The question arises, what is the corresponding best one dimensional approximation to w_1 and w_2, argued by the fact that $\dim \operatorname{span}\{w_1, w_2\} = 1$ if and only if q is an eigenvector. To this end, consider the optimality condition

$$\max_{\|z\|=1} \left(|z^* w_1|^2 + |z^* w_2|^2 \right) \tag{2.3}$$

to generate a projection (1.2) of the eigenvalue problem (1.1).

[5] We define the outer boundary to be the boundary of the smallest convex polytope containing the eigenvalues.

[6] Suppose q is an eigenvector of (1.1) such that q is orthogonal against Mq (and hence against Nq as well). Then q is a disastrous alternative for z.

Proposition 2.2 *Let $w_1, w_2 \in \mathbb{C}^n$ be non-orthogonal unit vectors. Then*

$$z = \frac{1}{\sqrt{2+2|w_1^* w_2|}}\left(\frac{w_1^* w_2}{|w_1^* w_2|} w_1 + w_2\right) \tag{2.4}$$

solves (2.3).

Proof Consider the matrix $\begin{bmatrix} w_1^* \\ w_2^* \end{bmatrix}$ as a linear operator from $\mathbb{C}^n$ to $\mathbb{C}^2$. Then solving (2.3) corresponds to finding the largest singular value of this matrix. The corresponding right singular vector yields z.

Observe that if w_1 and w_2 are linearly dependent, then $z = w_2$.

As opposed to $z = q$ of the Rayleigh quotient, with z given by (2.4) the aim is to predict, in terms of an optimality condition, what the common image of Nq and Mq in the Grassmannian $\mathrm{Gr}_1(\mathbb{C}^n)$ appears to be. Only thereafter, by inserting z into (1.2), an approximate eigenvalue is determined. This construction can be interpreted as an optimal approximation to the "partial" generalized Schur decomposition with respect to a single approximate eigenvector q. That is, in the generalized Schur decomposition

$$M = ZT_1Q^* \text{ and } N = ZT_2Q^* \tag{2.5}$$

with upper triangular T_1 and T_2 and unitary Z and Q. Now the vector q is an approximation to the first column of Q. In terms of this, our aim is to optimally predict what the first column of Z is.

Whenever $q^*N^*Mq \neq 0$, this choice of z yields us the quotient (1.3) which we call, because of the construction, optimal. If Nq and Mq are nonzero and orthogonal, then there is no arguable vector to pick and hence the quotient is not defined in a reasonable way.[7]

2.2 Field of optimal quotients

Collecting all the Rayleigh quotients yields the field of values. Analogously, collecting all the optimal quotients gives rise to the following notion.

Definition 2.2 Assume the matrix subspace (2.1) is nonsingular. The set

$$\left\{ \frac{q^*N^*Mq}{|q^*N^*Mq|} \frac{\|Mq\|}{\|Nq\|} \ : \ q^*N^*Mq \neq 0 \right\}$$

is said to be the field of optimal quotients of the eigenvalue problem (1.1) such that, additionally, if $Mq = 0$ (resp. $Nq = 0$) for $q \neq 0$, define the optimal quotient to have the value 0 (resp. ∞).[8] The field of optimal quotients is denoted by $\mathscr{F}(M,N)$.

[7] This is not a serious issue. If Nq and Mq are orthogonal, then q is an exceptionally poor approximation to an eigenvector. It is practically hopeless to produce any useful eigenvalue approximations with such a q.

[8] If $Mq = 0$ (resp. $Nq = 0$), choose z to be a unit vector in the direction of Nq (resp. Mq).

Consider (2.3). Inserting z given by (2.4) into $|z^*w_1|^2+|z^*w_2|^2$ yields

$$1+|w_1^*w_2| \tag{2.6}$$

which is a function of q alone, defined whenever $q^*N^*M^*q\neq 0$. The maximum value of this function is two, attained precisely at the eigenvectors of the eigenvalue problem (1.1) corresponding to the eigenvalues $\notin\{0,\infty\}$. Hence, this equivalently converts the eigenvalue problem into an optimization problem on the unit sphere of $\mathbb{C}^n$.

By construction, we have the following theorem.

Theorem 2.1 *The field of optimal quotients of the eigenvalue problem* (1.1) *contains its eigenvalues.*

For vectors nearby eigenvectors we have nearby eigenvalues as follows.

Proposition 2.3 *Suppose $q=v+\varepsilon$, where $v\in\mathbb{C}^n$ is an eigenvector corresponding to an eigenvalue $\lambda\neq\infty$ of the eigenvalue problem* (1.1)*. Then*

$$\frac{q^*N^*Mq}{|q^*N^*Mq|}\frac{\|Mq\|}{\|Nq\|}=\lambda+O(\|\varepsilon\|).$$

for $\varepsilon\in\mathbb{C}^n$ sufficiently small in norm.

Proof We have $Mv=\alpha w$ and $Nv=\beta w$ with a unit vector w and $\alpha,\beta\in\mathbb{C}$. Clearly $\lambda=\frac{\alpha}{\beta}$. If $\alpha=0$, then $\|Mq\|=O(\|\varepsilon\|)$ and the claim follows. So let us assume $\alpha\neq 0$. Then $(Nq)^*Mq=\alpha\overline{\beta}+\overline{\beta}w^*M\varepsilon+\alpha(N\varepsilon)^*w+(M\varepsilon)^*N\varepsilon=\alpha\overline{\beta}+O(\|\varepsilon\|)$. Thus, since $\alpha\overline{\beta}\neq 0$, by Taylor expanding we have $\frac{q^*N^*Mq}{|q^*N^*Mq|}=\frac{\alpha\overline{\beta}}{|\alpha\overline{\beta}|}+O(\|\varepsilon\|)$ for ε sufficiently small in norm. Clearly, $\frac{\alpha\overline{\beta}}{|\alpha\overline{\beta}|}$ has the same argument as $\frac{\alpha}{\beta}$. Similarly, $\frac{\|Mq\|}{\|Nq\|}=\frac{|\alpha|}{|\beta|}+O(\|\varepsilon\|)$, completing the proof.

Under stronger assumptions, $O(\|\varepsilon\|^2)$ estimates result; see Theorem 2.5 below.

Since our algorithms in Section 3 for approximating eigenpairs rely on using vectors (2.4) and respective optimal quotients, let us make some preliminary remarks on the location and properties of $\mathscr{F}(M,N)$.

Theorem 2.2 *There holds*

$$\mathscr{F}(N,M)=\frac{1}{\mathscr{F}(M,N)}$$

and

$$\mathscr{F}(e^{i\theta_1}M,e^{i\theta_2}N)=e^{i(\theta_1-\theta_2)}\mathscr{F}(M,N)$$

for $\theta_1,\theta_2\in\mathbb{R}$.

Proof The first claim follows from the identity

$$\frac{q^*M^*Nq}{|q^*M^*Nq|}\frac{\|Nq\|}{\|Mq\|}=\overline{\frac{q^*N^*Mq}{|q^*N^*Mq|}}\frac{1}{\frac{\|Mq\|}{\|Nq\|}}=\frac{1}{\frac{q^*N^*Mq}{|q^*N^*Mq|}\frac{\|Mq\|}{\|Nq\|}}$$

for any $q\in\mathbb{C}^n$. The latter claim is obvious.

The following observation is of particular significance.

Theorem 2.3 *Suppose $U \in \mathbb{C}^{n\times n}$ is unitary and $Y \in \mathbb{C}^{n\times n}$ invertible. Then*

$$\mathscr{F}(M,N) = \mathscr{F}(UMY^{-1}, UNY^{-1}).$$

Proof For any $q \in \mathbb{C}^n$, set $w = Yq$. Then we have

$$\frac{q^*N^*Mq}{|q^*N^*Mq|}\frac{\|Mq\|}{\|Nq\|} = \frac{w^*(NY^{-1})^*MY^{-1}w}{|w^*(NY^{-1})^*MY^{-1}w|}\frac{\|MY^{-1}w\|}{\|NY^{-1}w\|} =$$

$$\frac{w^*(UNY^{-1})^*UMY^{-1}w}{|w^*(UNY^{-1})^*UMY^{-1}w|}\frac{\|UMY^{-1}w\|}{\|UNY^{-1}w\|}$$

yielding the claim.

Typically, whenever realistic, generalized eigenvalue problems are transformed into standard problems. Because of the invariance under a change of basis in the domain, for the field of optimal quotients this takes place implicitly in the following way.

Corollary 2.1 *Suppose N is invertible. Then $\mathscr{F}(M,N) = \mathscr{F}(MN^{-1}, I)$.*

In the standard eigenvalue problem $N = I$. Then we have a spectral mapping theorem for the inversion as follows.

Corollary 2.2 *Suppose M is invertible. Then $\mathscr{F}(M^{-1}, I) = \frac{1}{\mathscr{F}(M,I)}$.*

Proof By Theorem 2.2 $\mathscr{F}(M,I) = \frac{1}{\mathscr{F}(I,M)}$ holds. Since $\mathscr{F}(I,M) = \mathscr{F}(M^{-1}, I)$ the claim follows.

We are not aware of any other inclusion sets for the eigenvalues satisfying a relationship like this under inversion. This means that, in general, $\mathscr{F}(M,I)$ differs from $\mathscr{F}(M)$, the field of values of the matrix M.

Example 2.1 Suppose $M \in \mathbb{C}^{n\times n}$ is unitary. Then $\mathscr{F}(M,I)$ is a subset of the unit circle and thereby a non-convex set in general. Recall that $\mathscr{F}(M)$ is always convex.

Example 2.2 In this small numerical experiment we had a normal $M \in \mathbb{C}^{100\times 100}$ while $N = I$. The eigenvalues of M were randomly generated. Then the field of values $\mathscr{F}(M)$ is well-understood by being a polygon with the vertices consisting of eigenvalues of M. (Take the convex hull of the eigenvalues.) Since M is invertible, we can expect by Corollary 2.2 that $\mathscr{F}(M,I)$ has hole in the middle. This can be seen in our numerical computations with a very preliminary algorithm. For the size of this hole, see Proposition 2.5.

Hence, the field of optimal quotients cannot be expected to be convex. Closedness cannot be assured either.

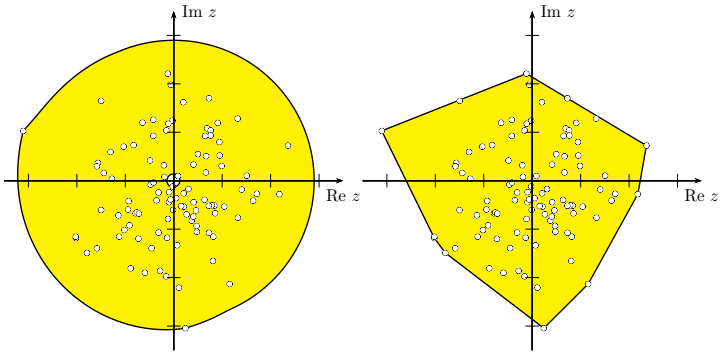

Fig. 2.1 Numerically computed $\mathscr{F}(M,I)$ and $\mathscr{F}(M)$ for Example 2.2. The eigenvalues of M are denoted by o's. Observe that $\mathscr{F}(M,I)$ is not simply connected.

Example 2.3 Consider the standard eigenvalue problem with $M=\begin{bmatrix}0&1\\0&0\end{bmatrix}$ and $N=I$. If $q=\begin{bmatrix}\alpha\\ \beta\end{bmatrix}$ is a unit vector, then (1.3) equals $\frac{\overline{\alpha}\beta}{|\overline{\alpha}\beta|}|\beta|$ whenever $\overline{\alpha}\beta\neq 0$. Moreover, $Mq=0$ for $\beta=0$. Consequently, $\mathscr{F}(M,N)$ equals the open unit disk. For comparison, $\mathscr{F}(M)$ is the closed disk of radius $\frac{1}{2}$ centered at the origin.

Closedness can be guaranteed in the following case.

Proposition 2.4 *If $0\notin\mathscr{F}(N^*M)$, then $\mathscr{F}(M,N)$ is closed.*

Proof Since $0\notin\mathscr{F}(N^*M)$, the map

$$q\longmapsto\frac{q^*N^*Mq}{|q^*N^*Mq|}\frac{\|Mq\|}{\|Nq\|}$$

is continuous. By the compactness of the set of unit vectors, the claim follows.

By Theorem 2.1, the matrix N^*M determines, through its field of values, the possible arguments of the eigenvalues of the eigenvalue problem (1.1). The following includes the standard Hermitian eigenvalue problem as a special case, allowing us to conclude when the eigenvalues are located on a line through the origin.

Theorem 2.4 *Suppose $N^*M=e^{i\theta}H$ for $\theta\in[0,2\pi)$ and a Hermitian matrix $H\in\mathbb{C}^{n\times n}$. Then $\mathscr{F}(M,N)$ is a subset of the line $\{re^{i\theta}\ :\ r\in\mathbb{R}\}$.*

For the location of the field of optimal quotients more generally, consider a closed annulus centered at the origin with the radii

$$r(M,N)=\min_{\|q\|=1}\frac{\|Mq\|}{\|Nq\|}\ \text{ and }\ R(M,N)=\max_{\|q\|=1}\frac{\|Mq\|}{\|Nq\|}. \tag{2.7}$$

If N is invertible, then $r(M,N)$ equals the smallest and $R(M,N)$ the largest singular value of the matrix MN^{-1}. Now the following inclusion region yields, in a sense, an analogue for $\mathscr{F}(M,N)$ of the Bendixson-Hirsch theorem [13, p.115].

Proposition 2.5 *$\mathscr{F}(M,N)$ is contained in that sector of the annulus centered at the origin with the radii* (2.7) *whose elements have the arguments of the field of values of N^*M.*

Proof If $\lambda = \frac{q^*N^*Mq}{|q^*N^*Mq|}\frac{\|Mq\|}{\|Nq\|}$, then it has the argument of q^*N^*Mq which, obviously, belongs to the field of values of N^*M. The modulus of λ is bounded from below by $r(M,N)$ and from above by $R(M,N)$. The exceptional case $Mq = 0$ (resp. $Nq = 0$) is covered similarly.

If N is singular, then there is an eigenvalue at infinity which is somewhat discomforting. To avoid this, solve the eigenvalue problem for some linear combinations (2.2) instead. Altogether, it is of interest to determine an annulus containing the eigenvalues which is tight in terms of the field of optimal quotients.

Example 2.4 Consider the standard eigenvalue problem with $N = I$ and M indefinite Hermitian and unitary. Thus, the spectrum of M consists of the points -1 and 1. Now $\mathscr{F}(M,N)$ is as tight as possible by equaling the spectrum of M. In particular, it is a non-convex set. If we take the linear combinations $A = M + bN$ with $b > 1$ and $B = N$, then $\mathscr{F}(A,B)$ equals the convex hull of the spectrum of A, i.e., the interval from $b-1$ to $2+b$.

For the tightest possible structure we have the following.

Example 2.5 Assume $Y \in \mathbb{C}^{n\times n}$ can be found such that MY^{-1} and NY^{-1} are linearly independent unitary matrices. By Theorem 2.3 and Example 2.1, then $\mathscr{F}(M,N)$ is a subset of the unit circle.

With the help of this example it is easy to show that, in general, $\mathscr{F}(M^*,N^*)$ does not equal $\mathscr{F}(M,N)$ conjugated.

Equivalence transformations are of major importance for the numerical solution of the eigenvalue problem. Matrix subspaces $\mathscr{V}$ and $\mathscr{W}$ are said to be equivalent if

$$\mathscr{W} = X\mathscr{V}Y^{-1} \tag{2.8}$$

holds for invertible matrices $X,Y \in \mathbb{C}^{n\times n}$. Recall that, by Theorem 2.3, the field of optimal quotients is preserved under an equivalence if X is unitary. Regarding diagonalizability, such a transformation is of importance by the fact that it yields the following improvement on Proposition 2.3.

Theorem 2.5 *In the eigenvalue problem* (1.1)*, assume UMY^{-1} and UNY^{-1} are diagonal for a unitary U and an invertible Y. Let $Mq_k = \lambda_k Nq_k$ with* $\dim\ker(M-\lambda_k N) = 1$. *If $\|q-q_k\| = \varepsilon$ and $\lambda_k \notin \{0,\infty\}$, then*

$$\left|\frac{q^*N^*Mq}{|q^*N^*Mq|}\frac{\|Mq\|}{\|Nq\|} - \lambda_k\right| = O(\varepsilon^2).$$

Proof Since the estimates are unitarily invariant, we may assume that $M = DY$ and $N = \hat{D}Y$ for diagonal matrices

$$D = \mathrm{diag}(d_1, d_2, \ldots, d_n) \text{ and } \hat{D} = \mathrm{diag}(\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_n).$$

Therefore Yq_k is a standard basis vector, let us say e_k. Since $\|q - q_k\| = \varepsilon$, we have $\|Y(q-q_k)\| \leq \|Y\|\varepsilon$. Therefore $Yq = ae_k + bv$ with a unit vector v orthogonal to e_k such that $a = 1 - c$ with $|c| \leq \|Y\|\varepsilon$ and $|b| \leq \|Y\|\varepsilon$.

Then

$$\frac{q^*N^*Mq}{|q^*N^*Mq|} = \frac{(ae_k + bv)^*\hat{D}^*D(ae_k + bv)}{|(ae_k + bv)^*\hat{D}^*D(ae_k + bv)|} = \tag{2.9}$$

$$\frac{\overline{\hat{d}_k}d_k|1-c|^2 + O(\varepsilon^2)}{|\overline{\hat{d}_k}d_k|1-c|^2 + O(\varepsilon^2)|} = \frac{\overline{\hat{d}_k}d_k}{|\overline{\hat{d}_k}d_k|} + O(\varepsilon^2) = \frac{\lambda_k}{|\lambda_k|} + O(\varepsilon^2)$$

and

$$\frac{\|Mq\|}{\|Nq\|} = \left(\frac{|d_k|^2|1-c|^2 + O(\varepsilon^2)}{|\hat{d}_k|^2|1-c|^2 + O(\varepsilon^2)}\right)^{\frac{1}{2}} = \frac{|d_k|}{|\hat{d}_k|} + O(\varepsilon^2) = |\lambda_k| + O(\varepsilon^2)$$

by using $\frac{1}{1+O(\varepsilon^2)} = 1 + O(\varepsilon^2)$ and $\left(1 + O(\varepsilon^2)\right)^{\frac{1}{2}} = 1 + O(\varepsilon^2)$. Then use $\frac{\overline{\hat{d}_k}d_k}{|\overline{\hat{d}_k}d_k|}\frac{|d_k|}{|\hat{d}_k|} = \lambda_k$ to have the claim.

The standard Hermitian eigenvalue problem is covered as a special case. However, the set of problems satisfying these assumptions is clearly much larger.

Non-unitary transformations X and Y appear in preconditioning large scale problems, i.e., when the aim is to steer the convergence of iterative methods for approximating eigenvalues to a particular part of the spectrum. (In practice this is always needed.) Then (1.1) gets transformed into an equivalent eigenvalue problem

$$XMY^{-1}x = \lambda XNY^{-1}x. \tag{2.10}$$

For a model problem on how this affects convergence, see Example 3.3 below. In particular, a non-unitary X can change the field of optimal quotients of the eigenvalue problem, i.e., possible approximations the optimal quotients can yield. Observe though that this is not a standard approach to preconditioning eigenvalue problems. (For standard approaches, see [30, Chapter 8] and [4] For a concise description, see [10, p. 57].) In an equivalence the spectrum remains intact. This is clearly positive. However, an equivalence changes any standard eigenvalue problem into a generalized one. This is not a serious issue since, to our mind, no distinction should be made between these two problems.

3 Optimal quotient iterations for the generalized eigenvalue problem

Optimal quotient iterations are derived relying either on exact or inexact inversion. Algorithm 1 is the most basic method and Algorithm 2 is supplemented with preconditioning. Algorithm 3 then aims to improve these by smartly taking linear combinations (2.2). When inexact inversion is used, an optimality construction is devised to guarantee a monotonic convergence behavior. This transforms Algorithms 1 and 2 into Algorithms 4 and 5.

Algorithm 1 Optimal quotient iteration for (1.1)

1: Read n-by-n matrices M and N and an approximate unit eigenvector q and a tolerance ε
2: **while** $\sigma_2([Mq\,Nq]) > \varepsilon$ **do**
3: Compute $w_1 = \frac{Mq}{\|Mq\|}$ and $w_2 = \frac{Nq}{\|Nq\|}$
4: Set $z = \frac{1}{\sqrt{2+2|w_1^* w_2|}}\left(\frac{w_1^* w_2}{|w_1^* w_2|} w_1 + w_2\right)$
5: Compute $l = \frac{q^* N^* Mq}{|q^* N^* Mq|} \frac{\|Mq\|}{\|Nq\|}$
6: Solve $(M - lN)\hat{q} = z$ and set $q = \hat{q}/\|\hat{q}\|$
7: **end while**

3.1 Optimal quotient iterations with accurate inversion

Given a unit vector $q \in \mathbb{C}^n$, the formula (2.4) yields a unit vector $z \in \mathbb{C}^n$ satisfying the optimality condition (2.3). To reverse the process, with z now available, it is a challenging problem to update q to have an improved approximation to an eigenvector of (1.1). As an extreme, suppose that q actually is an eigenvector of (1.1). Then it is not clear whether knowing $z = Mq/\|Mq\|$ alone is of any direct use.

If it is affordable to apply the inverse using, e.g., sparse direct solvers, then there is a natural operation on z. Since z is our best prediction of what the common one dimensional image of q under M and N appears to be, it can be used to update q. This means solving

$$(M - \lambda N)\hat{q} = z \tag{3.1}$$

with $\lambda = \frac{q^* N^* Mq}{|q^* N^* Mq|} \frac{\|Mq\|}{\|Nq\|}$. Thereafter set $q = \hat{q}/\|\hat{q}\|$. In particular, repeating this back and forth gives rise to an optimal quotient iteration; see Algorithm 1. There we denote by $\sigma_2([Mq\,Nq])$ the smallest singular value of the n-by-2 matrix $[Mq\,Nq]$, measuring how near the vectors Mq and Nq are to being linearly dependent. Of course, exact linear dependency corresponds to q being an eigenvector.[9]

We regard the generalized eigenvalue problem (1.1) de facto as a linear independency problem. Algorithm 1 transforms this into finding singularities of the resolvent operator

$$\lambda \longmapsto (M - \lambda N)^{-1}$$

by revealing its growth through the power method whenever (3.1) is solved. In particular, consider the standard eigenvalue problem, i.e., suppose $N = I$. Then Algorithm 1 notably differs from the Rayleigh quotient iteration by the fact that there holds $\frac{q^* Mq}{|q^* Mq|}\|Mq\| \neq q^* Mq$ as well as $z \neq q$ in general. (For the Rayleigh quotient iteration, see [25], [26, pp. 75–85] and [9, p. 457]. For the matters of implementation and software, see [14].)

Example 3.1 To compare Algorithm 1 against the Rayleigh quotient iteration, we took a standard eigenvalue problem with $M \in \mathbb{C}^{100\times 100}$ having eigenvalues circling the origin; see the right panel of Figure 3.1. We experimented by running twenty times Algorithm 1 versus the Rayleigh quotient iteration. Convergence with Algorithm 1 was distinctly faster.

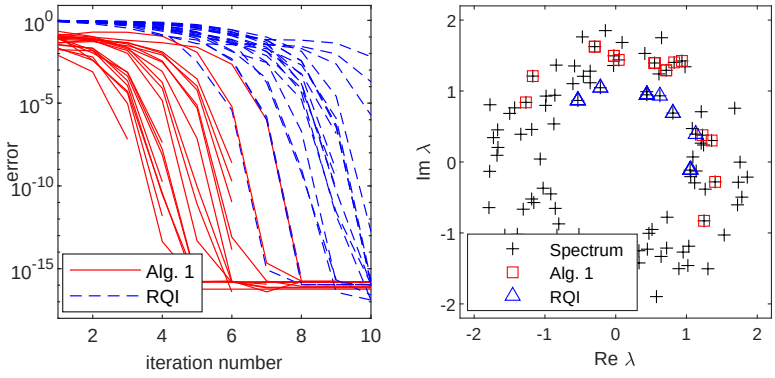

Fig. 3.1 For Example 3.1 convergence speeds with Algorithm 1 versus the Rayleigh quotient iteration. On the left panel we have convergence histories of the relative error $\frac{|\lambda_k-\lambda|}{|\lambda|}$ in the $\log_{10}$ scale. (Early termination means an eigenvalue has been found.) On the right panel we have the exact eigenvalues and the found eigenvalues.

Under additional assumptions we can expect the convergence of Algorithm 1 to be extraordinarily rapid, i.e., cubic.

Theorem 3.1 *Under the assumptions of Theorem 2.5, there holds* $\left\| \frac{\hat{q}}{\|\hat{q}\|} - q_k \right\| = O(\varepsilon^3)$, *where* $\hat{q}$ *denotes the solution of* (3.1).

Proof To have z of Proposition 2.2, we need

$$\frac{w_1^* w_2}{|w_1^* w_2|} = \frac{(Mq)^* Nq}{|(Mq)^* Nq|} = \frac{\overline{\lambda_k}}{|\lambda_k|} + O(\varepsilon^2)$$

by using (2.9). Then

$$z = \frac{1}{\sqrt{2+2|w_1^2 w_2|}} \left(\frac{\overline{\lambda_k}}{|\lambda_k|} \frac{\frac{ad_k}{|ad_k|} e_k + \mathbf{O}(\varepsilon_\perp)}{(1+O(\varepsilon^2))^{1/2}} + \frac{\frac{a\hat{d}_k}{|a\hat{d}_k|} e_k + \mathbf{O}(\varepsilon_\perp)}{(1+O(\varepsilon^2))^{1/2}} + \mathbf{O}(\varepsilon^2) \right)$$

with boldfaced $\mathbf{O}(\varepsilon_\perp)$ denoting vectors orthogonal to e_k. This yields

$$z = const.(e_k + \mathbf{O}(\varepsilon_\perp) + \mathbf{O}(\varepsilon^2)).$$

Then, by using Theorem 2.5, solving for $\hat{q}$ yields

$$(M - (\lambda_k + O(\varepsilon^2))N)^{-1} z = const.Y^{-1}(\frac{1+O(\varepsilon^2)}{O(\varepsilon^2)} e_k + \mathbf{O}(\varepsilon_\perp)) =$$

$$const.\frac{1}{O(\varepsilon^2)}(Y^{-1} e_k + \mathbf{O}(\varepsilon_\perp^3)).$$

[9] Prof. B. Parlett suggests using the optimal quotient as soon as $|q^* N^* Mq| \geq 0.95\|Mq\|\|Nq\|$ [24].

$j=0$	$j=1$	$j=2$
5.06...	5.21413...	5.21431974337712...
5	5.2131...	5.21431974318...

Table 3.1 For Example 3.2, on the second row the performance of Algorithm 1 and on the third row the Rayleigh quotient iteration. Here j denotes the number of linear system solves. The corresponding numerically computed eigenvalue is $\lambda = 5.21431974337753...$ (Computed with Matlab.)

From this it follows (after possibly multiplying $\hat{q}$ by $e^{i\theta}$ for $\theta \in \mathbb{R}$) that $\| \frac{\hat{q}}{\|\hat{q}\|} - q_k \| = O(\varepsilon^3)$.

Suppose one is interested in eigenvalues only (and not in eigenvectors). If the assumptions of Theorem 2.5 hold for the adjoint of $\mathscr{V}$, i.e., for $\mathrm{span}\{M^*, N^*\}$, then eigenvalues should be approximated with help of the adjoint. Then, of course, by conjugating one obtains approximations to eigenvalues of the original problem (1.1).

For bench-marking, it is certainly of interest to compare Algorithm 1 against the Rayleigh quotient iteration in the case of standard Hermitian eigenvalue problem. (By Theorem 2.4, the approximations are then real.) As is well known, then the Rayleigh quotient iteration also converges cubically; see [25, 30].

Example 3.2 To see how Algorithm 1 fares against the Rayleigh quotient iteration, let us take the tiny but educative Hermitian example carefully treated in [32, Example 27.1]. That is

$$M = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 4 \end{bmatrix}$$

and $N = I$ while the starting vector is $q = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T / \sqrt{3}$. Algorithm 1 provides more accurate approximations by yielding three correct digits more after two inversion steps; see Table 3.1. This is particularly startling by the fact that we approximate Hermitian problems "non-symmetrically" by choosing $z \neq q$.

This is actually no accident. Both of these iterations rely on the power method combined with applying the inverse. Better results with Algorithm 1 are, aside from using $z \neq q$, due to considerably greedier approximations towards the dominating eigenvalue with the optimal quotient (1.3) compared with the Rayleigh quotient. This can be made quantitative in terms of the following proposition.[10]

Proposition 3.1 *Assume M is Hermitian and $N = I$. Then the optimal quotient* (1.3) *equals $\|Mq\|$ for $q^*Mq > 0$ and $-\|Mq\|$ for $q^*Mq < 0$.*

It is noteworthy that, because of this difference, the same (not so good) starting vector can converge to different eigenvalues depending on whether Algorithm 1 or the Rayleigh quotient iteration is executed; see Example 3.1. Hence the difference is genuine not just in terms of the speed.

[10] For a Hermitian matrix M with the dominating eigenvalue λ_1 holds $|q^*Mq| \leq \|Mq\| \leq \|M\| = |\lambda_1|$ for any unit vector q. The inequalities are strict unless q is an eigenvector related with λ_1.

Algorithm 2 Optimal quotient iteration for (1.1) with an initial guess l

1: Read n-by-n matrices M and N and an approximate unit eigenvector q and a tolerance ε. Set $X = (M-lN)^{-1}$
2: **while** $\sigma_2([Mq\ Nq]) > \varepsilon$ **do**
3: Compute $l = \frac{q^*N^*X^*XMq}{|q^*N^*X^*XMq|}\frac{\|XMq\|}{\|XNq\|}$
4: Form $\hat{q} = (M-lN)^{-1}(M+lN)q$ and set $q = \hat{q}/\|\hat{q}\|$
5: Set $X = (M-lN)^{-1}$
6: **end while**

In practice the convergence needs to be steered to a desired part of the spectrum which means we have to modify Algorithm 1 accordingly. This is achieved by preconditioning the eigenvalue problem by performing equivalence transformations (2.10). To this end, consider the following simple model problem.

Example 3.3 Take two real diagonal matrices $M = \mathrm{diag}(a,b)$ and $N = \mathrm{diag}(c,d)$ such that

$$\frac{a}{c} \approx 1 \text{ and } \frac{b}{d} > 1.$$

Suppose we are interested in locating the eigenvalue near one, i.e., the smaller in magnitude. This forces $ac > bd$ but we want to have a better understanding of the relative sizes of the entries. To simplify computations, let the starting vector be equally supported, i.e., $q = \left[\begin{array}{cc} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{array}\right]^*$. (This is a fairly realistic assumption.) Then in the first iteration of Algorithm 1 we have $l = \sqrt{\frac{a^2+b^2}{c^2+d^2}} \approx \frac{a}{c}(1+\frac{b^2}{2a^2})(1-\frac{d^2}{2c^2})$ assuming $a > b$ and $c > d$. Then $M - lN \approx \mathrm{diag}(\frac{b^2c^2-a^2d^2}{2ac^2}, b - \frac{ad}{c})$. This is guaranteed to yield a small $(1,1)$-entry if $a \gg b^2$ and $c \gg d^2$. If this is not the case, then this can be achieved by preconditioning (2.10) with X being an approximation to the inverse of $M-N$. Also Y can be taken to be an approximation to $M-N$.

Indeed, consider (2.10) by taking $X = (M-l_0N)^{-1}$ for some $l_0 \in \mathbb{C}$ nearby which we look for eigenvalues. We consider only preconditioning from the left, so let us put $Y = I$. Then Algorithm 1 yields after one iteration $\hat{q} = const.\,(M-lN)^{-1}(M+lN)q$, where $l = \frac{q^*N^*X^*XMq}{|q^*N^*X^*XMq|}\frac{\|XMq\|}{\|XNq\|}$. This can be repeated by using l computed in forming $X = (M-lN)^{-1}$, to dynamically change the preconditioner; see Algorithm 2. Then the evolution of the approximate eigenvector can be elegantly expressed in terms of consecutive applications of "generalized" Cayley transformations of very particular type. That is, after k iterations the approximate eigenvector is given by

$$(M-l_kN)^{-1}(M+l_kN)\cdots(M-l_2N)^{-1}(M+l_2N)(M-l_1N)^{-1}(M+l_1N)q, \quad (3.2)$$

where l_j are the respective optimal quotients. Hence we take k consecutive applications of Cayley transformations.

There is one additional ingredient to speed-up iterations. That is, taking linear combinations (2.2) can improve convergence. To weight the effects of inverse iteration, at each step consider

$$(M-l_jN)x = \lambda Bx$$

Algorithm 3 Optimal quotient iteration for (1.1) with an initial guess l

1: Read n-by-n matrices M and N and an approximate unit eigenvector q and a tolerance ε. Set $X = (M - lN)^{-1}$
2: **while** $\sigma_2([Mq\ Nq]) > \varepsilon$ **do**
3: Compute $k = \frac{(XNq)^*q}{|(XNq)^*q|}\frac{\|q\|}{\|XNq\|}$
4: Form $\hat{q} = (M-(k+l)N)^{-1}(M+(k-l)N)q$ and set $q = \hat{q}/\|\hat{q}\|$
5: Set $l = k+l$
6: Set $X = (M-lN)^{-1}$
7: **end while**

with the preconditioning $X = (M - l_j N)^{-1}$. Here B is some linear combination of M and N which should be as well-conditioned as possible; see [16, Section 4.1]. If finding such a B is, e.g., too costly, simply put $B = N$. This is done in Algorithm 3.

3.2 Optimal quotient iterations with approximate inversion

It is not realistic to apply the inverse very accurately. That is, for very large problems one almost certainly needs to invoke iterative methods for solving linear systems (3.1) with an acceptable number of flops. Also each appearance of X requires solving a linear system. It is very likely that at this point preconditioning becomes necessary. Conceptually this makes preconditioning eigenproblems simple.

This is not a trivial matter, though. In practice (3.1) can be expected to be indefinite, i.e., not easy to precondition successfully. Therefore iterative methods may encounter notable difficulties such that solving these linear systems to the full machine precision poses, in general, an insurmountable challenge. So the question arises, how to proceed when only approximate solving is affordable?

We suggest the following. Suppose having somehow generated

$$Q_k = \begin{bmatrix} q_1 & q_2 & \cdots & q_k \end{bmatrix} \in \mathbb{C}^{n\times k}$$

with orthonormal columns. Then, to have a best approximate eigenvector, find the least linearly independent element for the eigenvalue problem (1.1) from the column space of Q_k. That is, from the span of the columns of Q_k find

$$\max_{v\in\mathbb{C}^k,\ \|v\|=1} \frac{|v^*\hat{N}^*\hat{M}v|^2}{\|\hat{N}v\|^2\|\hat{M}v\|^2}, \tag{3.3}$$

where $\hat{M} = MQ_k$ and $\hat{N} = NQ_k$ are of size n-by-k with $k \ll n$. Solving this is of particular significance for the numerical solution of eigenvalue problems. See (2.6) in particular. Numerical methods to solve (3.3) are devised in Section 5.

Definition 3.1 The set

$$\left\{ \frac{v^*\hat{N}^*\hat{M}v}{\|\hat{N}v\|\|\hat{M}v\|} \ :\ v\in\mathbb{C}^k,\ \|v\| = 1 \right\}$$

is said to be the scaled field of values of the pair $\hat{M},\hat{N} \in \mathbb{C}^{n\times k}$ corresponding to $Q_k \in \mathbb{C}^{n\times k}$ with orthonormal columns.

The scaled field of values belongs to the unit disk. For a qualitative interpretation, the points closest to the unit circle correspond to best approximate eigenvectors available from the span of the columns of Q_k.

Proposition 3.2 *Under the assumptions of Definition 3.1, let $\hat{M} = Q_1R_1$ and $\hat{N} = Q_2R_2$ be the QR decompositions of $\hat{M}$ and $\hat{N}$ which are assumed to be of rank k both. Then* (3.3) *is bounded above by the largest singular value of $Q_2^*Q_1$ squared.*

Proof We have

$$\frac{|v^*\hat{N}^*\hat{M}v|^2}{\|\hat{N}v\|^2\|\hat{M}v\|^2} = \frac{|(R_2v)^*Q_2^*Q_1R_1v|^2}{\|R_2v\|^2\|R_1v\|^2} \leq \max_{\|u\|=1,\,\|v\|=1} |v^*Q_2^*Q_1u|^2$$

proving the claim.

This means that the optimal quotient (3.3) is bounded above by one which is attained if and only if the span of the columns of Q_k contains an eigenvector corresponding to an eigenvalue $\notin \{0,\infty\}$.

Example 3.4 Assume M and N are unitary. Then the scaled field of values is simply $\mathscr{F}(\hat{N}^*\hat{M})$. Clearly, $\hat{N}^*\hat{M}$ is a contraction.

Example 3.5 For the standard eigenvalue problem we have $N = I$, so that we are concerned with

$$\left\{ \frac{v^*Q_k^*MQ_kv}{\|MQ_kv\|} \;:\; v \in \mathbb{C}^k,\ \|v\| = 1 \right\}$$

then.

To turn this into an algorithm, suppose v solves (3.3). This yields a maximal linear dependency method for approximating eigenvalues in the sense that $q = Q_kv$ can be regarded as a best eigenvector approximation available from the span of the columns of Q_k. With this vector q, take one step of Algorithm 1 or 2, possibly combined with approximate solving of the appearing linear systems. With the updated q, after orthonormalizing, augment Q_k with this new column vector. Repeating this transforms Algorithm 1 into Algorithm 4. Similarly, Algorithm 2 becomes Algorithm 5.

Algorithms 4 and 5 perform optimally as follows. When we augment Q_k with a column q_{k+1} to have

$$Q_{k+1} = \begin{bmatrix} q_1 & q_2 & \cdots & q_k & q_{k+1} \end{bmatrix} \in \mathbb{C}^{n\times(k+1)}, \tag{3.4}$$

then with this the corresponding maximum (3.3) obviously cannot decrease. Thereby optimal eigenvector approximations are necessarily generated yielding a monotonic convergence behavior with respect to (3.3). Thereby, methodologically, this belongs to the same category as the GMRES method for solving linear systems.

In practice $k+1$, the number of stored columns in (3.4) cannot get too large. Like in solving large and sparse linear systems with algorithms such as the GMRES method, restarting provides a way to save storage. Then only the so far best approximate eigenvector q is kept stored and the process is restarted with $Q_1 = \begin{bmatrix} q \end{bmatrix}$.

Algorithm 4 Maximal linear dependency method for (1.1)

Read n-by-n matrices M and N and an approximate unit eigenvector q
Set $Q = [q]$
for $j = 1, 2, \ldots$ **do**
    Compute $w_1 = \frac{Mq}{\|Mq\|}$ and $w_2 = \frac{Nq}{\|Nq\|}$
    Set $z = \frac{1}{\sqrt{2+2|w_1^* w_2|}} \left(\frac{w_1^* w_2}{|w_1^* w_2|} w_1 + w_2 \right)$
    Compute $l = \frac{q^* N^* M q}{|q^* N^* M q|} \frac{\|Mq\|}{\|Nq\|}$
    Possibly approximately, solve $(M - lN)\hat{q} = z$
    Orthonormalize $\hat{q}$ against the columns of Q to have q
    Set $Q = [Q\, q]$
    Solve (3.3) and set $q = Qv$
end for

Algorithm 5 Maximal linear dependency method for (1.1) with an initial guess l

Read n-by-n matrices M and N and an approximate unit eigenvector q and a tolerance ε.
Set $Q = [q]$
while $\sigma_2([Mq\, Nq]) > \varepsilon$ **do**
    Possibly approximately, solve $(M - lN)y_1 = Mq$ and $(M - lN)y_2 = Nq$
    Set $w_1 = \frac{y_1}{\|y_1\|}$ and $w_2 = \frac{y_2}{\|y_2\|}$
    Compute $l = \frac{w_2^* w_1}{|w_2^* w_1|} \frac{\|y_1\|}{\|y_2\|}$
    Possibly approximately, solve $(M - lN)\hat{q} = (M + lN)q$
    Orthonormalize $\hat{q}$ against the columns of Q to have q
    Set $Q = [Q\, q]$
    Solve (3.3) and set $q = Qv$
end while

One should be aware of the following. Suppose $\frac{|v^* \hat{N}^* \hat{M} v|^2}{\|\hat{N}v\|^2 \|\hat{M}v\|^2} = 1 - \varepsilon$ with $\varepsilon > 0$. Then for the amount of linear dependency of the corresponding vectors, consider the matrix

$$\left[\begin{array}{cc} \frac{\hat{M}v}{\|\hat{M}v\|} & \frac{\hat{N}v}{\|\hat{N}v\|} \end{array} \right]. \tag{3.5}$$

Its second singular value is $\sqrt{\varepsilon}$. Hence inspecting (3.3) leads to a loss of accuracy when looking at how small the respective residual vector is. This means, in particular, that when ε is near the machine precision, then no improved accuracy can be expected with Algorithms 4, 5. Should this not suffice, we recommend executing Algorithm 1 thereon by using the approximate eigenvector computed as an initial guess. This strategy is used in the numerical experiments in Section 5.

4 Optimal projection of the eigenvalue problem

So far we have been concerned with single vector projections (1.2) and respective iterations. Single vector methods are fundamental since they typically pave way to more general methods involving subspaces. This is of importance since subspaces provide more options for restarting and hence ways to affect and speed up the convergence; see [5] and [7] and references therein. Subspaces arise also in model reduction so it is of interest to cleverly choose them more generally. Regarding the Rayleigh quotient,

see [25, Chapter 11.3] how the extension is done for the standard Hermitian eigenvalue problem. In the non-Hermitian case one deals with the Arnoldi method [30]. For the finite section (or Galerkin) method in general, see [1] and references therein.

The notion of optimal quotient also has a natural extension to involve subspaces. To this end, consider projecting the eigenvalue problem assuming $Q_k \in \mathbb{C}^{n\times k}$ with orthonormal columns has been generated. The task is now to find in some sense optimal $Z_k \in \mathbb{C}^{n\times k}$ with orthonormal columns so as to have

$$Z_k^* M Q_k = \lambda Z_k^* N Q_k, \tag{4.1}$$

i.e., an optimally projected eigenvalue problem of size k-by-k. With this one aims to approximate the following structure.

Definition 4.1 Subspaces $\mathscr{Q}, \mathscr{Z} \subset \mathbb{C}^n$ are said to provide an invariant structure for the eigenvalue problem (1.1) if

$$(M-\lambda N)\mathscr{Q} = \mathscr{Z}$$

for all but finite $\lambda \in \mathbb{C}$.

Clearly, $\mathscr{Q}$ and $\mathscr{Z}$ are necessarily of the same dimension. The one dimensional case corresponds to a solution of the eigenvalue problem (1.1).

Invariance corresponds to having (partially) solved the eigenvalue problem. In lack of this, by considering the column spaces, regard

$$MQ_k \text{ and } NQ_k \tag{4.2}$$

as elements of the Grassmannian $\mathrm{Gr}_k(\mathbb{C}^n)$ of k dimensional subspaces of $\mathbb{C}^n$. Denote by $\hat{Z}_k$ and $\tilde{Z}_k$ matrices having orthonormal columns spanning the column spaces of (4.2). Analogously to the one dimensional case, consider finding their best approximation from $\mathrm{Gr}_k(\mathbb{C}^n)$. By using the Frobenius norm, then a multidimensional version of the optimality condition (2.3) can be formulated as

$$\max_{Z_k\in\mathrm{Gr}_k(\mathbb{C}^n)} \left(\|Z_k^*\hat{Z}_k\|_F^2 + \|Z_k^*\tilde{Z}_k\|_F^2 \right), \tag{4.3}$$

where $Z_k \in \mathrm{Gr}_k(\mathbb{C}^n)$ is represented by an n-by-k matrix having orthonormal columns. For the eigenvalue problem this means finding a partial equivalence transformation which maximizes the attainable projection onto the column spaces of (4.2). The task appears approximation theoretically natural and is well-defined in $\mathrm{Gr}_k(\mathbb{C}^n)$, i.e., it does not depend on the choice of the representing matrices $\hat{Z}_k$, $\tilde{Z}_k$ and Z_k.

To solve (4.3), let us assume that $\hat{Z}_k$ and $\tilde{Z}_k$ have been chosen in such a way that $\hat{Z}_k^*\tilde{Z}_k = \Sigma$ is diagonal with non-negative entries. More precisely, compute the SVD

$$\hat{Z}_k^*\tilde{Z}_k = U\Sigma V^* \tag{4.4}$$

of $\hat{Z}_k^*\tilde{Z}_k$ with unitary $U,V \in \mathbb{C}^{k\times k}$ and a diagonal matrix Σ with non-negative entries. Then take $\hat{Z}_k U$ and $\tilde{Z}_k V$ to replace the original $\hat{Z}_k$ and $\tilde{Z}_k$. Denote now by $\hat{z}_j$ and $\tilde{z}_j$

Algorithm 6 Computing Z_k satisfying (4.3)

1: Read n-by-n matrices M and N and n-by-k matrix Q_k with orthonormal columns
2: Compute the QR factorizations $MQ_k = Q_1R_1$ and $NQ_k = Q_2R_2$
3: Compute the SVD $Q_1^*Q_2 = U\Sigma V^*$
4: Set $\hat{Z} = Q_1U$ and $\tilde{Z} = Q_2V$
5: With the columns $\hat{z}_j$ and $\tilde{z}_j$ of $\hat{Z}$ and $\tilde{Z}$
6: **for** $l = 1,\ldots,k$ **do**
7: Compute $a = \hat{z}_j^*\tilde{z}_j$ and $b = \arg a$
8: Set $z_j = (e^{ib}\hat{z}_j + \tilde{z}_j)/\sqrt{2+2|a|}$
9: **end for**
10: Set $Z_k = [z_1 \cdots z_k]$

the columns of $\hat{Z}_k$ and $\tilde{Z}_k$. With these vectors as w_1 and w_2, form the columns of Z_k according to Proposition 2.2. This yields

$$\|Z_k^*\hat{Z}_k\|_F^2 + \|Z_k^*\tilde{Z}_k\|_F^2 = \sum_{j=1}^{k} 1+\sigma_j, \tag{4.5}$$

where σ_j are the diagonal entries of Σ. Hence, the value of (4.5) is bounded by $2k$ such that if $2k$ is attained, then one has computed an invariant structure for the eigenvalue problem (1.1).

Theorem 4.1 *Assume $\hat{Z}_k, \tilde{Z}_k \in \mathbb{C}^{n\times k}$ have orthonormal columns. Then a solution Z_k to* (4.3) *satisfies* (4.5).

Proof By dropping indices, consider the linear map

$$Z \mapsto \begin{bmatrix}\hat{Z}^*\\ \tilde{Z}^*\end{bmatrix} Z = \begin{bmatrix}\hat{Z}^*Z\\ \tilde{Z}^*Z\end{bmatrix} \tag{4.6}$$

from $\mathbb{C}^{n\times k}$ to $\mathbb{C}^{2k\times k}$. By the construction in connection with (4.4), we may assume that $\hat{Z}$ and $\tilde{Z}$ are such that $\hat{Z}^*\tilde{Z} = \Sigma$ is diagonal with non-negative entries. By regarding (4.6) as acting columnwise on Z, in terms of the Kronecker product we may consider

$$\operatorname{vec}(Z) \mapsto M\operatorname{vec}(Z) = \begin{bmatrix}I\otimes\hat{Z}^*\\ I\otimes\tilde{Z}^*\end{bmatrix}\operatorname{vec}(Z) \tag{4.7}$$

with the identity I being of size k-by-k, so that M is of size $2k^2$-by-nk. We have

$$MM^* = \begin{bmatrix}I\otimes\hat{Z}^*\\ I\otimes\tilde{Z}^*\end{bmatrix}\begin{bmatrix}I\otimes\hat{Z} & I\otimes\tilde{Z}\end{bmatrix} = \begin{bmatrix}I\otimes I & I\otimes\Sigma\\ I\otimes\Sigma & I\otimes I\end{bmatrix}.$$

Hence, the nonzero singular values of this map are determined by taking k times the positive square root of the eigenvalues of $\begin{bmatrix}I & \Sigma\\ \Sigma & I\end{bmatrix}$. These eigenvalues are $1\pm\sigma_j$ for $j = 1,\ldots,k$. Because of (4.7), the problem separates and becomes that of how to position k orthonormal vectors with respect to the $2k$ singular values $\sqrt{1\pm\sigma_j}$ for $j = 1,\ldots,k$ for the matrix $\begin{bmatrix}\hat{Z}^*\\ \tilde{Z}^*\end{bmatrix}$. This is the dual problem of approximating with the singular value decomposition with the optimal solution satisfying (4.5).

To see how this works in the most familiar case, consider projecting the standard eigenvalue problem, i.e., assume having $N = I$. Then the Arnoldi method constitutes a basic scheme to produce approximations for large scale problems [30,31]. The idea is, in a nutshell, to better exploit the information provided by the power method combined with extending the notion of Rayleigh quotient to involve subspaces. This means that Q_k results from orthonormalizing the sequence

$$q, Mq, M^2q, \ldots, M^{k-1}q \tag{4.8}$$

by executing the Arnoldi method with a starting vector $q \in \mathbb{C}^n$; see, e.g., [30, Chapter 6]. From this point on our optimal scheme proceeds differently as follows. Clearly, $\tilde{Z}_k = Q_k$ holds. Then $\hat{Z}_k$ is obtained by orthonormalizing the columns of $\tilde{Z}_{k+1}\tilde{H}_k$, where

$$M\tilde{Z}_k = \tilde{Z}_{k+1}\tilde{H}_k \tag{4.9}$$

with $\tilde{H}_k \in \mathbb{C}^{(k+1)\times k}$ being of upper-Hessenberg type. By using $\hat{Z}_k$ and $\tilde{Z}_k$, compute Z_k to satisfy (4.5). Because of the construction, the columns of Z_k are linear combinations of the columns of $\hat{Z}_{k+1}$ and thereby we are dealing with a Krylov subspace method.[11] It is noteworthy that the projected eigenvalue problem is now likely to be generalized. This very much underscores that it is artificial to make any distinction between standard and generalized eigenvalue problems. The following numerical example illustrates this.

Example 4.1 A Markov model of a random walk on a triangular grid [30, Section 2.5.1] is a well-documented test for basic iterative eigensolvers; see [30, Example 4.1] and [30, Example 6.1].[12] The problem is standard such that the eigenvalues at the right end of the spectrum are of interest. Here the Matlab script of [30, p. 44] is used to generate $M \in \mathbb{R}^{n\times n}$ while $N = I$. We took $n = 5050$. The starting vector was randn$(n,1)$ divided by its norm. In Figure 4 we have compared the classical Arnoldi method against the optimal Arnoldi method, drawn vertically while the iteration number runs horizontally. Whenever the optimal Arnoldi method yields an extreme Ritz value appearing as a pair (i.e., two genuinely complex extreme Ritz values), the approximation is of the same order as given by the classical Arnoldi method. Whenever the optimal Arnoldi method yields is single real extreme Ritz value, the approximation is better than that given by the classical Arnoldi method.

5 Numerical experiments

In what follows, numerical experiments on an eigenvalue problem in magnetohydrodynamics are presented to illustrate the convergence and computational cost of the algorithms devised in Section 3. For Algorithms 4 and 5. we first need to devise a

[11] By a Krylov subspace method in meant that all the basis vectors computed are expressible by polynomials in M of degree at most the number of iterates applied to q.

[12] Also to be found from the matrix market at http://math.nist.gov/MatrixMarket/

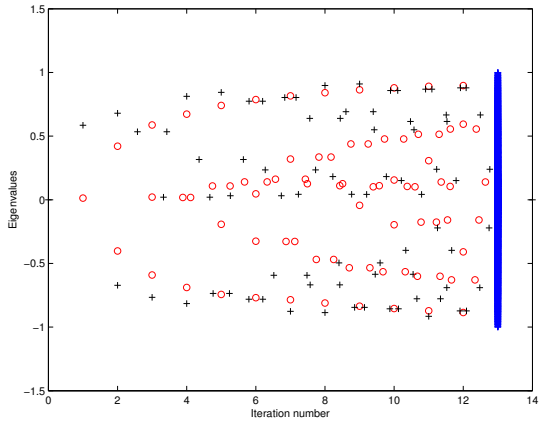

Fig. 4.1 For Example 4.1, the convergence of the classical Arnoldi method (depicted with 'o') and the optimal Arnoldi method (depicted with '+'), drawn vertically. The first 12 steps are shown. Numerically computed eigenvalues of M are all real, drawn vertically on the right.

method to solve the optimization problem (3.3). Regarding computational complexity, this constitutes a critical part of the iteration, aside form solving (3.1).

To numerically solve (3.3), consider minimizing the function

$$f(v) = -\frac{|v^*\hat{N}^*\hat{M}v|^2}{\|\hat{N}v\|^2\|\hat{M}v\|^2},$$

where $\hat{M} \in \mathbb{C}^{n\times k}$, $\hat{N} \in \mathbb{C}^{n\times k}$ and $v \in \mathbb{C}^k$. Relaxing the constraint $\|v\| = 1$ allows using descend methods as well as the Newton method. This is certainly attractive. However, then there is no decrease in "radial directions", a fact which must be taken into account in devising a method to minimize f.

In practice the number of columns in $\hat{M}$ and $\hat{N}$ is very small compared with n. Then the additional cost of computing the Hessian matrix in the Newton method is not excessive. The indefiniteness of the Hessian calls for special attention, though, as does its singularity in the proximity of the local minima. These issues will be addressed after investigating the actual calculation of the gradient and the Hessian of f.

To differentiate the real valued function f with respect to complex variable v, it is convenient to use the so-called Wirtinger calculus; see the highly accessible presentation [17]. Regarding f for a moment as a function of two separate vectors v and $\overline{v}$, the gradient with respect to v is

$$\begin{aligned} g_v(v) &= \left(\frac{\partial}{\partial v} f(v)\right)^* \\ &= -\frac{(v^*\hat{N}^*\hat{M}v)\hat{M}^*\hat{N}v + (v^*\hat{M}^*\hat{N}v)\hat{N}^*\hat{M}v}{\|\hat{N}v\|^2\|\hat{M}v\|^2} - f(v)\left(\frac{\hat{N}^*\hat{N}v}{\|\hat{N}v\|^2} + \frac{\hat{M}^*\hat{M}v}{\|\hat{M}v\|^2}\right) \end{aligned} \tag{5.1}$$

and, since $f(v)$ is real valued, the gradient with respect to $\overline{v}$ is $g_{\overline{v}}(v) = \overline{g_v(v)}$. The second order derivatives with respect to v and $\overline{v}$ are, accordingly,

$$\begin{aligned}
\mathscr{H}_{vv} &= \frac{\partial}{\partial v}\left(\frac{\partial f}{\partial v}\right)^* \\
&= -\frac{\hat{M}^*\hat{N}vv^*\hat{N}^*\hat{M} + v^*\hat{N}^*\hat{M}v\hat{M}^*\hat{N}}{\|\hat{N}v\|^2\|\hat{M}v\|^2} - \frac{\hat{N}^*\hat{M}vv^*\hat{M}^*\hat{N} + v^*\hat{M}^*\hat{N}v\hat{N}^*\hat{M}}{\|\hat{N}v\|^2\|\hat{M}v\|^2} \\
&+ \frac{(v^*\hat{N}^*\hat{M}v)\hat{M}^*\hat{N}v + (v^*\hat{M}^*\hat{N}v)\hat{N}^*\hat{M}v}{\|\hat{N}v\|^2\|\hat{M}v\|^2}\left(\frac{v^*\hat{N}^*\hat{N}}{\|\hat{N}v\|^2} + \frac{v^*\hat{M}^*\hat{M}}{\|\hat{M}v\|^2}\right) \\
&- \left(\frac{\hat{N}^*\hat{N}v}{\|\hat{N}v\|^2} + \frac{\hat{M}^*\hat{M}v}{\|\hat{M}v\|^2}\right)[g(v)]^* \\
&- f(v)\left(\frac{\hat{N}^*\hat{N}}{\|\hat{N}v\|^2} - \frac{\hat{N}^*\hat{N}vv^*\hat{N}^*\hat{N}}{\|\hat{N}v\|^4} + \frac{\hat{M}^*\hat{M}}{\|\hat{M}v\|^2} - \frac{\hat{M}^*\hat{M}vv^*\hat{M}^*\hat{M}}{\|\hat{M}v\|^4}\right) \qquad (5.2)
\end{aligned}$$

$$\begin{aligned}
\mathscr{H}_{\overline{v}v} &= \frac{\partial}{\partial \overline{v}}\left(\frac{\partial f}{\partial v}\right)^* = -\frac{\hat{M}^*\hat{N}vv^T\hat{M}^T\overline{\hat{N}}}{\|\hat{N}v\|^2\|\hat{M}v\|^2} - \frac{\hat{N}^*\hat{M}vv^T\hat{N}^T\overline{\hat{M}}}{\|\hat{N}v\|^2\|\hat{M}v\|^2} \\
&+ \frac{(v^*\hat{N}^*\hat{M}v)\hat{M}^*\hat{N}v + (v^*\hat{M}^*\hat{N}v)\hat{N}^*\hat{M}v}{\|\hat{N}v\|^2\|\hat{M}v\|^2}\left(\frac{v^T\hat{N}^T\overline{\hat{N}}}{\|\hat{N}v\|^2} + \frac{v^T\hat{M}^T\overline{\hat{M}}}{\|\hat{M}v\|^2}\right) \\
&- \left(\frac{\hat{N}^*\hat{N}v}{\|\hat{N}v\|^2} + \frac{\hat{M}^*\hat{M}v}{\|\hat{M}v\|^2}\right)[g(v)]^T + f(v)\left(\frac{\hat{N}^*\hat{N}vv^T\hat{N}^T\overline{\hat{N}}}{\|\hat{N}v\|^4} + \frac{\hat{M}^*\hat{M}vv^T\hat{M}^T\overline{\hat{M}}}{\|\hat{M}v\|^4}\right) \qquad (5.3)
\end{aligned}$$

and $\mathscr{H}_{v\overline{v}} = \overline{\mathscr{H}_{\overline{v}v}}$, $\mathscr{H}_{\overline{v}\overline{v}} = \overline{\mathscr{H}_{vv}}$. Using this notation, the Taylor expansion of $f(v+\Delta v)$ with respect to $\left[\begin{smallmatrix} v \\ \overline{v}\end{smallmatrix}\right]$ around v up to second order is given by

$$f(v+\Delta v) \approx T_2(v+\Delta v) = f(v) + \begin{bmatrix} g_v^* & \overline{g_v}^* \end{bmatrix}\begin{bmatrix}\Delta v \\ \overline{\Delta v}\end{bmatrix} + \frac{1}{2}\begin{bmatrix}\Delta v^* & \overline{\Delta v}^*\end{bmatrix}\begin{bmatrix}\mathscr{H}_{vv} & \mathscr{H}_{v\overline{v}} \\ \mathscr{H}_{\overline{v}v} & \mathscr{H}_{\overline{v}\overline{v}}\end{bmatrix}\begin{bmatrix}\Delta v \\ \overline{\Delta v}\end{bmatrix}.$$

To transform the problem from $\mathbb{C}^k$ to $\mathbb{R}^{2k}$, denote $z_r = \left[\begin{smallmatrix}\operatorname{Re} z \\ \operatorname{Im} z\end{smallmatrix}\right]$ for any $z \in \mathbb{C}^k$, and note that

$$\begin{bmatrix} z \\ \overline{z}\end{bmatrix} = Jz_r, \quad J = \begin{bmatrix} I & iI \\ I & -iI\end{bmatrix}.$$

With somewhat abusive convention $f(v) = f(v_r)$, the above Taylor expansion reads

$$T_2(v_r + \Delta v_r) = f(v_r) + g_{v_r}(v)^T\Delta v_r + \frac{1}{2}\Delta v_r^T\mathscr{H}_{v_r}\Delta v_r,$$

where the real gradient g_{v_r} and the real Hessian $\mathscr{H}_{v_r}$ are readily obtained as

$$g_{v_r} = 2\begin{bmatrix}\operatorname{Re} g_v \\ \operatorname{Im} g_v\end{bmatrix}, \qquad \mathscr{H}_{v_r} = \begin{bmatrix} I & iI \\ I & -iI\end{bmatrix}^*\begin{bmatrix}\mathscr{H}_{vv} & \mathscr{H}_{\overline{v}v} \\ \mathscr{H}_{\overline{v}v} & \mathscr{H}_{vv}\end{bmatrix}\begin{bmatrix} I & iI \\ I & -iI\end{bmatrix}.$$

Newton's method The Newton direction $p = -\mathscr{H}_{v_r}^{-1} g_{v_r}$ is guaranteed to be a descend direction whenever the Hessian $\mathscr{H}_{v_r}$ is positive definite. Because this is not to be expected, the method needs to be modified. Keeping in mind that the size of the Newton iteration equation is small, an eigendecomposition $\mathscr{H}_{v_r} = \sum_{j=1}^{2k} d_j x_j x_j^*$ is computationally affordable, where x_j denotes an eigenvector corresponding to the eigenvalue d_j. Assuming that the eigenvalues are in ascending order, we can make the following observations.

First, for $\varepsilon > 0$ a positive definite modification of $\mathscr{H}_{v_r}$ is available by

$$\widetilde{\mathscr{H}_{v_r}} = \sum_{j=1}^{2k} \tilde{d}_j x_j x_j^* \qquad \text{with } \tilde{d}_j = \begin{cases} d_j, & d_j > \varepsilon \\ \max\{\varepsilon, -d_j\}, & d_j \le \varepsilon. \end{cases}$$

(For this and other modifications see, e.g., [8].) This modification guarantees that the Newton direction $p = -\sum_{j=1}^{2k} \tilde{d}_j^{-1}(g_{v_r}, x_j)x_j$ is a descend direction. If the unit step of Newton method satisfies the so-called strong Wolfe conditions, it should be accepted; otherwise a line search in the direction of p is to be conducted [23].

Second, if the eigenvalues $d_1, \ldots, d_s$ are negative, a positive linear combination d of the corresponding eigenvectors gives a direction of negative curvature. A line search in this direction may then result in more descent than the Newton direction. One can try to predict this by comparing, say, unit-step descent in both directions [11].

Finally, the radial direction relates to two eigenvalues closest to zero in absolute value, say $d_{j^\dagger}$ and $d_{j^\star}$. Then one can remove this direction by setting

$$p = -\sum_{j \neq j^\dagger, j^\star} \tilde{d}_j^{-1}(g_{v_r}, x_j)x_j.$$

Collecting these conditions yields Algorithm 7.

This allows us to perform numerical experiments with Algorithms 1 and 4. (We always monitor the relative error $\log_{10} \frac{|\lambda_k - \lambda|}{\lambda}$.)

Example 5.1 The performance of Algorithms 1 and 4 was tested using data for calculation of Alfven spectra in magnetohydrodynamics from the NEP collection [2]. Real M and N of size 4800×4800 are nonsymmetric and positive definite, respectively. A random unit starting vector q was used. The computations were conducted with MATLAB (version R2016a), and the iterates were compared to the nearest eigenvalue obtained by MATLAB's built-in solver for eigenproblems. For the line search, the algorithm of Moré and Thuente as implemented by Sandia National Laboratories in Poblano Toolbox v. 1.0 was used. The progress of the iterations are shown in Figure 5.1. The iteration converged to $-2035.6 - i6283.3$, which is the second largest eigenvalue in modulus. Observe that once Algorithm 4 attains accuracy $\sqrt{\varepsilon}$ with ε near the machine precision, we switch to Algorithm 1; see the discussion in connection with (3.5).

In Example 5.1, the iterations tended to converge to some "isolated" eigenvalues. To steer the iteration to a specific eigenvalue near an initial guess, choose either Algorithms 2, 3 or 5. In the following example we experiment with Algorithms 2 and 5.

Algorithm 7 Minimization of $f(v)$.

Choose positive quantities ε_1, ε_2 and ε_3 with $\varepsilon_2 < \varepsilon_3$.
Append $\hat{M}$ by Mq, $\hat{N}$ by Nq, and v by zero.
Set $f_{\text{old}} = -1$ and calculate $f(v)$.
while $|f(v) - f_{\text{old}}| > \varepsilon_1$, **do**
    Calculate g_{v_r} and $\mathscr{H}_{v_r}$.
    Find the eigendecomposition $\mathscr{H}_{v_r} = \sum_{j=1}^{2k} d_j x_j x_j^*$ with $d_1 \le d_2 \le \cdots \le d_{2k}$.
    Set $\tilde{d}_j = \begin{cases} d_j, & d_j > \varepsilon_2 \\ \max\{\varepsilon_2, -d_j\}, & d_j \le \varepsilon_2. \end{cases}$
    Calculate the Newton direction $p = -\sum_{j \neq j^\dagger, j^\star} \tilde{d}_j^{-1}(g_{v_r}, x_j) x_j$.
    if $d_1 < -\varepsilon_3$ **then**
        Form a direction d of negative curvature.
        if $T_2(v + p/\|p\|) < T_2(v + d/\|d\|)$ **then**
            Perform a line search in direction p starting with step size 1.
        else
            Perform a line search in direction d.
        end if
    else
        Perform a line search in direction p starting with step size 1.
    end if
    Set $f_{\text{old}} = f(v)$ and calculate new $f(v)$.
end while

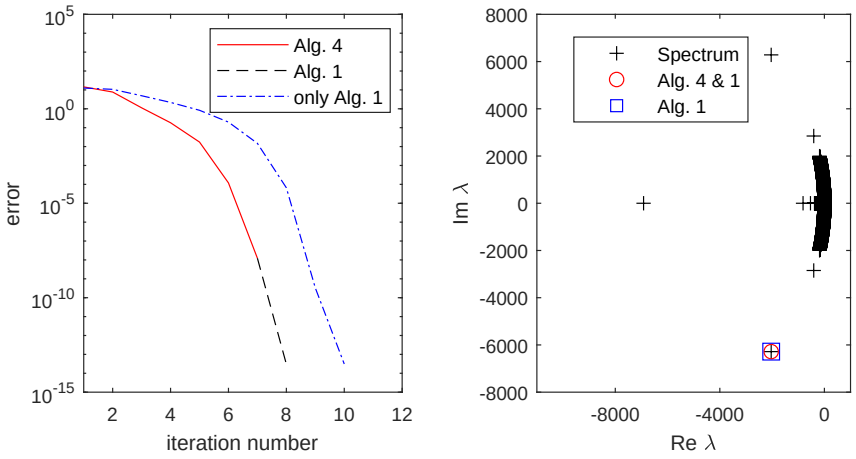

Fig. 5.1 For Example 5.1, on the left the relative error produced with a combination of Algorithm 4 and 1 compared against Algorithm 1. The eigenvalues of the Alfven spectrum problem of size 4800×4800 are depicted on the right together with the eigenvalue of convergence.

Example 5.2 To calculate eigenvalues in the top branch of the Alfven spectrum, an initial guess $\lambda = -175 + i1965$ was made in Algorithms 2 and 5. To statistically study the convergence, a set of 20 normally distributed unit starting vectors was generated. Algorithm 2 converged to $\lambda_1 = -180.8 + i1963.3$ on all but one occasion, on which it converged to $\lambda_2 = -181.0 + i1961.7$. Algorithm 5 converged to λ_1 on four occasions and to $\lambda_3 = -180.6 + i1964.9$ on three occasions. On thirteen occasions it converged

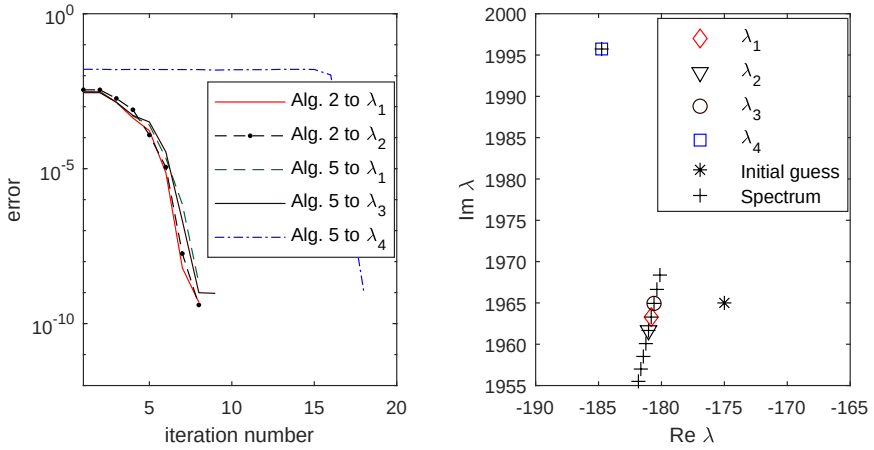

Fig. 5.2 For Example 5.2, on the left the relative error of Algorithms 2 and 5 with initial guess $-175+i1965$. Depending on the starting vector, the convergence rate of Algorithm 5 is comparable to Algorithm 2, or it may converge to a different eigenvalue. The figure on the right provides a closer look into the Alfven spectrum in the area of interest together with computed approximations.

to $\lambda_4 = -184.8 + i1995.7$. In the cases of λ_1 and λ_3, Algorithm 5 performed like Algorithm 2. In the case of λ_4, the convergence was slow until the algorithm started finding an eigenvalue farther away, after which the decay of the error was rapid. Convergence histories representing the different cases are depicted in Figure 5.2.

Finite precision analysis is beyond the scope of this paper. However, based on numerical experiments, Algorithm 1 appears to yield numerically the most accurate results, attaining the relative error of the order of machine precision in case of convergence. It may be advisable, by using an appropriate switching criterion, to switch to Algorithm 1 after the convergence appears to start accelerate.

Example 5.3 To compare Algorithm 2 with Algorithm 3, another set of 20 normally distributed unit starting vectors were generated. With Algorithm 3 $\sigma_2([Mq\ Nq])$ attained the machine precision consistently. Algorithm 2 was run until $\sigma_2([Mq\ Nq])$ reached the square root of the machine precision and then switching was performed so that Algorithm 1 was executed thereon by using the best eigenvector approximation computed so far as an initial guess. A Rayleigh quotient variant for generalized eigenvalue problems [26, Chapter 15] was run to compare the performance.

With Algorithm 2 we converged to λ_1 in 17 occasions and to λ_3 in three occasion. With Algorithm 3 we converged to λ_1 in 18 occasions and to λ_2 in 2 occasions. The Rayleigh quotient iteration converged to λ_1 in 16 occasions and to λ_3 in 4 occasions. For all three methods, the loss of linear independency of Mq and Nq, measured in terms of $\sigma_2([Mq\ Nq])$, consistently reached the same magnitude as $\sigma_2([M\tilde{q}\ N\tilde{q}])$, where $\tilde{q}$ was computed with MATLABs eigs-function (which was also used as a reference). On the average, Algorithm 3 was about 1-2 iterations faster than Algorithms 2 and 1 combined. Algorithms 2 and 1 combined was, in turn, about one iteration faster than the Rayleigh quotient iteration. See the left panel of Figure 5.3 for a typical case.

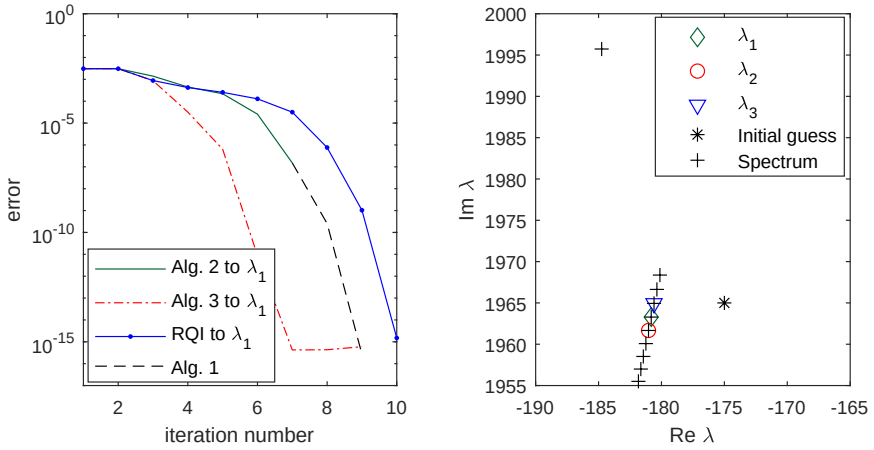

Fig. 5.3 For Example 5.3, on the left typical convergence histories of Algorithms 2 and 3 compared against the Rayleigh quotient iteration. The reference eigenvalue, generated by MATLAB's `eigs`-function, was found within the machine precision. The corresponding eigenvalues and approximations are depicted on the right.

In very large eigenvalue problems applying the inverse of $M - lN$ in these algorithms may only be feasible by executing iterative methods. (The inverse appears in these algorithms either in solving (3.1) or in applying X.) Then the linear systems cannot be expected to be solved to the full machine precision. The accuracy reached is determined by the quality of preconditioning. The following example is meant to illustrate how this affects the convergence of iterations.

Example 5.4 In this experiment Algorithms 2 and 5 were modified by replacing each appearance of a linear system with a preconditioned iterative solver. The respective linear systems were solved until the relative residual attained the tolerance 10^{-6}. This was done by executing the GMRES method using the row-sum modified Crout version of ILU as a preconditioner with the droptolerance 10^{-10}.[13] The outer iteration was terminated if $\sigma_2([Mq\ Nq])$ reached $\varepsilon = 2^{-26}$, the square root of the machine precision, or if it ceased to decrease. A set of 20 starting vectors were generated with 10 ones at random positions, divided by the norm. In all but one occasions the algorithms yielded convergence to the same eigenvalue with the relative error varying between 10^{-6} and 10^{-9}. In these occasions, Algorithm 5 reached the relative error 10^{-6} approximately two iterations earlier on the average. In Figure 5.4, a typical convergence history for both algorithms using the same starting vector is presented. In this case Algorithm 5 seems to be more robust to errors in computations.

[13] We do not claim this to be a realistic preconditioning strategy.

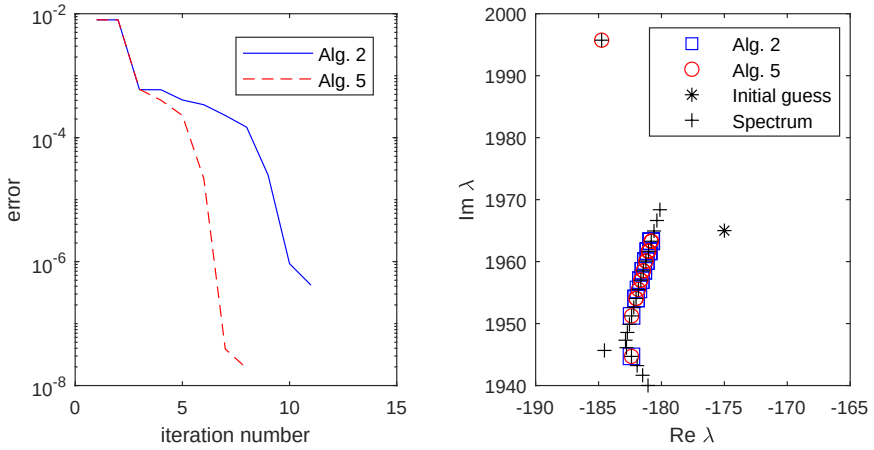

Fig. 5.4 For Example 5.4, on the left an illustration of the convergence histories for Algorithms 2 and 5 when the linear systems are approximately solved with the GMRES method. On the right the spread of convergence points with the initial guess $\lambda = -175 + i1965$.

Conclusions

A notion of quotient was introduced for the eigenvalue problem (1.1), be it generalized or not. A criterion was formulated to compute the quotient by optimally approximating the associated vectors Mq and Nq for $q \in \mathbb{C}^n$ given. (The Rayleigh quotient, on the contrary, is based on optimally approximating an eigenvalue, i.e., a scalar.) The choice gives rise to a spectral inclusion set which appears to have strikingly different properties from the field of values. The construction extends to subspaces and allows, for example, optimally performing model reduction.

The optimal quotient and the associated projector vector z were used in deriving an iterative method for approximating eigenpairs. This method can be argued to yield more accurate results than the Rayleigh quotient iteration, without any additional cost. At the limit a partial generalized Schur decomposition corresponding to a single eigenvector is attained. All in all, the set-up is such that there is no need to transform generalized eigenvalue problems into standard problems. This means, in particular, that replacing exact inversions with inaccurate solvers causes no problems. This is due to an optimality criterion for extending subspaces which guarantees improved approximations at every step.

In terms of equivalence transformations, preconditioning can be incorporated into this set-up in a natural way. This allows steering the convergence to desired parts of the spectrum. Numerical experiments were performed to illustrate all these aspects of the method proposed.

References

1. Arveson, W.: $\mathbf{C}^*$-algebras and numerical linear algebra. J. Funct. Anal. **122**, 333–360 (1994)
2. Bai, Z., Day, D., Demmel, J., Dongarra, J.: Non-Hermitian Eigenvalue Problem (NEP) Collection. Available: `http://math.nist.gov/MatrixMarket/data/NEP/mhd/mhd.html`
3. Bathe, K.-J.: The subspace iteration method - Revisited. Computers and Structures **126**, 177–183 (2013)
4. Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H. (eds.): Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide. SIAM, Philadelphia (2000)
5. Beattie, C.A., Embree, M., Sorensen, D.C.: Convergence of polynomial restart Krylov methods for eigenvalue computations. SIAM Rev. **47**, 492–515 (2005)
6. Einstein, E., Johnson, C.R., Lins, B., Spitkovsky, I.: The ratio field of values. Linear Algebra Appl. **434**, 1119–1136 (2011)
7. Fokkema, D. Sleijpen, G., van der Vorst, H.: Jacobi-Davidson style QR and QZ algorithms for the reduction of matrix pencils. SIAM J. Sci. Comput. **20**, 94–125 (1998)
8. Gill, P.E., Murray, W., Wright, M.H.: Practical Optimization. Academic Press, London (1981)
9. Golub, G.H., van Loan, C.F.: Matrix Computations (4th ed.). The Johns Hopkins University Press, Baltimore (2013)
10. Golub, G.H., van der Vorst, H.: Eigenvalue computation in the 20th century. J. Comput. Appl. Math. **123**, 35–65 (2000)
11. Gould, N.I.M., Lucidi, S., Roma, M., Toint, P.L.: Exploiting negative curvature directions in linesearch methods for nonconstrained optimization. Optim. Methods Softw. **14**, 75–98 (2000)
12. Gustafson, K.E., Rao, D.K.M.: Numerical Range. Springer Verlag, New York (1997)
13. Halmos, P.: A Hilbert Space Problem Book (Graduate Texts in Mathematics, Volume 19). Springer-Verlag, New York (1982)
14. Hernández V., Román J.E., Thomás A., Vidal V.: Single vector iteration methods in SLEPc, SLEPc Technical Report STR-2, (2005)
15. Horn, R.A., Johnson, C.R.: Topics in Matrix Analysis. Cambridge Univ. Press, Cambridge (1991)
16. Huhtanen M., Seiskari O.: Matrix intersection problems for conditioning. Banach Center Publ. **112**, 195–210 (2017)
17. Kreutz-Delgado, K.: The complex gradient operator, and the CR-Calculus. (2009) `http://arxiv.org/abs/0906.4835v1` [math.OC]
18. Lancaster, P., Rodman, L.: Canonical forms for Hermitian matrix pairs under strict equivalence and congruence. SIAM Rev. **47**, 407–443 (2005)
19. Lehoucq, R.B., Meerbergen, K.: Using the generalized Cayley transformations within an inexact Krylov sequence method. SIAM J. Matrix Anal. Appl. **20**, 131–148 (1998)
20. Mawhin J.: Spectra in mathematics and in Physics: from the dispersion of light to nonlinear eigenvalues. CIM bulletin, Centro Internacional de Mathematica Portugal, **29**, 3–13 (2011)
21. Martin, M.: On different definitions of numerical range. J. Math. Anal. Appl. **433**, 877–866 (2016)
22. Moré, J.J. , Thuente, D.J.: Line search algorithms with guaranteed sufficient decrease. ACM Transactions on Mathematical Software **20**, 286–307 (1994)
23. Nocedal, J., Wright,S.J.: Numerical Optimization (2nd ed.). Springer, New York (2006)
24. Parlett, B.: a private communication, (2018)
25. Parlett, B.: The Rayleigh quotient iteration and some generalizations for nonnormal matrices. Math. Comp. **28**, 679–693 (1974)
26. Parlett, B.: The Symmetric Eigenvalue Problem. Classics in Applied Mathematics 20, SIAM, Philadelphia (1997)
27. Psarrakos, P.J.: Numerical range of linear pencils. Linear Algebra Appl. **317**, 127–141 (2000)
28. Ruhe, A.: Rational Krylov algorithms for nonsymmetric eigenvalue problems. In: Recent advances in iterative methods, IMA Vol. Math. Appl., vol. 60, pp. 149–164. Springer, New York (1994)
29. Ruhe, A.: Rational Krylov: a practical algorithm for large sparse nonsymmetric matrix pencils. SIAM J. Sci. Comput. **19**, 1535–1551 (1998)
30. Saad, Y.: Numerical Methods for Large Eigenvalue Problems, 2nd Edition. SIAM, Philadelphia (2011)
31. Sorensen, D.C.: Numerical methods for large eigenvalue problems. In: Acta Numerica, pp. 519–584. Cambridge University Press, Cambridge (2002)
32. Trefethen, L.N., Bau, D. III: Numerical Linear Algebra. SIAM, Philadelphia (1997)