Numer. Algorithms manuscript No.
(will be inserted by the editor)

An Approximate Eigensolver for Self-Consistent Field Calculations

Harald Hofstätter · Othmar Koch

July 19, 2013

Abstract In this paper, we give a comprehensive error analysis for an approximate solution method for the generalized eigenvalue problems arising for instance in the context of electronic structure computations based on density functional theory. The solution method has been demonstrated to excel as compared to established solvers in both computational effort and scaling for parallelization. Here we estimate the improvement provided by our proposed subspace method starting from the initial approximations for instance provided in the course of the self-consistent field iteration, showing that in general the approximation quality is improved by our method to yield sufficiently accurate eigenvalues.

Keywords Electronic structure computations; Density functional theory; Generalized eigenvalue problem; Iterative diagonalization

Mathematics Subject Classification (2000) 65F15,65F08,65Z05

1 Introduction

In this paper, we discuss an approximate numerical solution method for the generalized eigenvalue problems which arise in each step of the fixed point iteration (commonly referred to as the *self-consistent field (SCF) cycle*) employed for the solution of the *Kohn–Sham equations* [10] of *density functional theory* [8] in large scale electronic structure computations [15, 19, 20].

Harald Hofstätter
Institute for Analysis and Scientific Computing, Vienna University of Technology
Wiedner Hauptstrasse 8–10, A-1040 Wien, Austria
E-mail: hofi@harald-hofstaetter.at

Othmar Koch
Institute for Analysis and Scientific Computing, Vienna University of Technology
Wiedner Hauptstrasse 8–10, A-1040 Wien, Austria
E-mail: othmar@othmar-koch.org

After discretization by *augmented plane waves plus local orbitals* [3,4,16] this implies the solution of large generalized eigenproblems with symmetric matrices. For most applications not all eigenvalues need to be computed, but commonly only the lowest 3–10% of these are needed. This suggests the use of an iterative subspace method for the approximation of the eigenvalues and eigenvectors, which only computes a fraction of the eigensystem which is of practical relevance. The algorithm discussed here was implemented in the WIEN2k code [4] and found to outperform previous solution methods in both computational effort and parallelization, see [2].

The iterative method for the diagonalization of generalized eigenproblems implemented formerly in the WIEN2k code was a blocked version of the Davidson method [5,6] which was introduced in [21]. Iterative methods for the problem at hand are also discussed in [1,11–13,17,24,27]:

In [27], the method of *RMM-DIIS (residual minimization/direct inversion in the iterative subspace*) is proposed and compared with the *Davidson* and *Block Davidson* methods. The latter has the disadvantage that the doubling in the dimension of the search space is prohibitive for large initial subspaces. Therefore the RMM-DIIS method is claimed to have the advantage that only matrices of the size of the number of previous iteration steps are necessary. However, in its original version the method is fundamentally sequential in nature which the authors recognize as a major drawback [27], and which in the light of the development of parallel and grid-enabled versions of the WIEN2k code makes this approximate diagonalization unattractive. Recently, a reformulation of RMM-DIIS [18] has brought this method into the scope for a parallel implementation, however. Another interesting approach was put forward in [26] where preconditioners similar to ours (based on approximations to the inverse of $(H - \lambda S)$) were tested. However, these methods are designed for sparse matrices.

A comparison with several other methods shows that (disregarding computational cost) the block Davidson method displays the best improvement in accuracy per iteration step due to the doubling of the search space [27]. Our aim is to avoid this doubling of the subspace.

Ref. [24] gives an overview of the state of the art of iterative diagonalization at that time, and demonstrates that a new *preconditioned conjugate gradient method* compares most favorably with *conjugate gradients*, *steepest descent* and *imaginary time propagation*. The VASP code [13], which is a highly efficient plane wave pseudopotential code, uses the RMM-DIIS method of [27] in a variant proposed in [17]. They claim this method to be superior for very large problems [11] if an unblocked, band-by-band iteration is used.

In more recent work, other subspace methods are put forward. [29] gives a method where subspace doubling is required only in the first step of the iterative solution and where parallelization is based on a decomposition of the physical domain. Since our method has a more general scope and only uses one iteration step, we do not consider this alternative. In [7] a variant of the power method realized in a subspace is introduced which does not offer the safeguard of working in a larger subspace and whose parallelization is not discussed.

In the present paper, we consider a blocked subspace method which is motivated by the code structure of WIEN2k and parallelization issues [2]. Both aspects suggest to refrain from a sequential "band–by–band" computation.

Our approximate diagonalization is motivated by the fact that the Davidson method previously implemented in the WIEN2k code [21] was recognized as unsatisfactory when the basis set was changed from the standard LAPW to the APW+lo basis set [16]. Apparently, the underlying discretization, the importance of non-diagonal terms (the local orbital contribution to the plane wave basis) and the adaptive basis set (the basis set changes slightly in the course of the SCF cycle) renders the preconditioning with only the diagonal elements $\mathrm{diag}^{-1}[H - \lambda S]$ inefficient. Our new method is motivated by the improvements promised by the Jacobi-Davidson method [22, 23, 25] as compared to the original Davidson method [6]. However, application of the subspace expansion from the Jacobi-Davidson method seems prohibitively expensive, hence we propose a simplification which uses an approximate computation of the subspace expansion related to an iterative solution of the associated linear system of equations. Furthermore, it was demonstrated in [2] that our method is superior to full diagonalization in efficiency and scales very well in a parallel implementation. We stress that the success of the method is linked to the structure of the problem considered. While the accuracy and efficiency is excellent for the problems solved in the WIEN2k code [2], we do not claim that it will excel for problems from other applications and thus does not necessarily represent a general purpose method. However, the present paper gives a general statement on the error behavior in Theorem 1 which is independent of the application problem.

In the course of the SCF iteration, good initial guesses are available for the approximation of the eigensystem of the generalized eigenvalue problems, since the problem data only changes moderately in the course of the iteration. Thus, in each iteration step it is sufficient to improve the numerical solution to an extent such that the required accuracy is achieved. In this paper, we are going to estimate the factor by which the approximation is improved by the update defined by our method. Numerical experiments show that indeed the bounds are sharp.

The outline of the paper is as follows: We introduce our subspace method to improve an initial approximation to the solution of a generalized eigenvalue problem in Section 2. In Section 3 we give the results of our error analysis, which estimate the factor by which the error of the initial approximation is reduced by applying one step of our method. Section 4 gives numerical experiments, showing that our error bounds are sharp and that the results also pertain to eigenvalue problems from real life applications. Finally, Appendix A contains the technical proof details of our main theorem from Section 3.

2 The Approximation Algorithm

We want to compute approximations to eigenvectors corresponding to the m lowest eigenvalues of the generalized eigenproblem

$$HX = SX\Lambda, \tag{1}$$

where the *Hamiltonian matrix* $H \in \mathbb{C}^{n\times n}$ is Hermitian (but not necessarily positive definite), the *overlap matrix* $S \in \mathbb{C}^{n\times n}$ is Hermitian and positive definite, and Λ is a diagonal matrix containing the (real!) eigenvalues. First, we specify the algorithm employed in the computation of the eigensystem. We consider only the real case ($H, S \in \mathbb{R}^{n\times n}$ symmetric) for simplicity. The adaptation of the algorithm and the analysis for the complex case is straightforward.

- Input:
$$Y = [y_1, \ldots, y_m] \in \mathbb{R}^{n\times m}.$$
Usually these are approximations to eigenvectors which were computed in the last SCF cycle.
- Compute the Ritz values (Rayleigh quotients)
$$\vartheta_j = \frac{y_j^T H y_j}{y_j^T S y_j}, \qquad j = 1, \ldots, m. \tag{2}$$
- Set up the search space $[Y\ Z] \in \mathbb{R}^{n\times 2m}$ with
$$z_j = H^{-1}(H - \vartheta_j S)y_j, \qquad j = 1, \ldots, m. \tag{3}$$
- Set up the reduced problem
$$\tilde{H}V = \tilde{S}V\Gamma, \tag{4}$$
where
$$\tilde{H} = [Y\ Z]^T H[Y\ Z] = \begin{bmatrix} Y^T HY & Y^T HZ \\ Z^T HY & Z^T HZ \end{bmatrix} \tag{5}$$
and
$$\tilde{S} = [Y\ Z]^T S[Y\ Z] = \begin{bmatrix} Y^T SY & Y^T SZ \\ Z^T SY & Z^T SZ \end{bmatrix}. \tag{6}$$
- Compute eigenvectors $V_{1:m}$ of (4) corresponding to the lowest m eigenvalues $\gamma_1 \leq \cdots \leq \gamma_m$ using, e.g., appropriate routines from LAPACK. We may assume that $V_{1:m}$ is orthonormal with respect to $\tilde{S}$, i.e., $V_{1:m}^T \tilde{S} V_{1:m} = I_m$.
- Compute new approximations
$$Y_{\text{new}} = [Y\ Z]V_{1:m} \tag{7}$$
to the eigenvectors of (1).

For practical computations, we have to take the possibility into account that in (4) the matrices $\tilde{H}$, $\tilde{S}$ are (nearly) singular, so that the reduced eigenproblem (4) admits no unique solution. For example this happens if some of the initial guesses y_j (almost) coincide with the corresponding exact eigenvectors of (1), so that the corresponding vectors z_j are (almost) zero. Our work-around is that if $z_j \approx 0$ (a condition which is of course easy to check) we simply delete the corresponding columns y_j, z_j in Y and Z, respectively (and take y_j as the computed result for the j-th eigenvector). It is very unlikely that the set of columns of $[YZ]$ is linearly dependent other than by the vanishing of some column z_j. We exclude the pathological cases in our analysis and will always assume that $[YZ]$ has full rank $2m$. Furthermore, we omit the case of multiple or vanishing eigenvalues for simplicity of the analysis.

The algorithm described above is designed to replace the *Davidson method* [6] employed previously [21], where (3) is replaced by

$$z_j = \mathrm{diag}^{-1}(H - \vartheta_j S)(H - \vartheta_j S)y_j, \quad j = 1, \ldots, m. \tag{8}$$

3 The Main Result

Let $X \in \mathbb{R}^{n\times n}$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_1 < \cdots < \lambda_n$ be an exact solution of (1). Due to the structure of (1), the columns of X are orthogonal with respect to S, and without restriction of generality we may assume

$$X^T S X = I_n, \tag{9}$$

where I_n denotes the $n \times n$ identity matrix. Clearly for (3) to be well-defined, we have to assume that H is non-singular, or equivalently that all eigenvalues λ_j are non-zero. For the analysis we further assume that all "discarded" eigenvalues are greater than zero,

$$0 \ < \lambda_{m+1} < \cdots < \lambda_n,$$

and that the moduli of the remaining eigenvalues are smaller than λ_{m+1},

$$-\lambda_{m+1} < \lambda_1 < \cdots < \lambda_m < \lambda_{m+1}.$$

These properties can always be enforced by changing the system matrix $H \to H+\gamma S$, with suitable $\gamma \in \mathbb{R}$ (this shifts the eigenvalues, but leaves the eigenvectors unaltered). Note that if all eigenvalues λ_j are distinct, then X is uniquely determined up to multiplication from the right by a diagonal matrix with entries ± 1. At one point in the following analysis we will have to assume in addition to the eigenvalues being distinct that these are even well separated. More precisely, if $O(\varepsilon)$ is the order of the error ΔY of the given approximate solution

$$Y = X_{1:m} + \Delta Y, \tag{10}$$

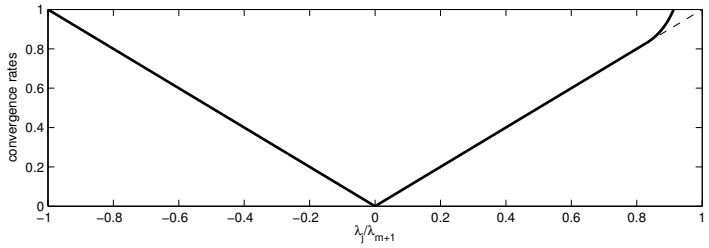

Fig. 1 Bounds for the convergence rates $\|\Delta y_{\mathrm{new},j}\|_{X^TS}/\|\Delta y_j\|_{X^TS}$ as given by Theorem 1, eqs. (14) (solid line) and (15) (for the special case $m = 1$, dashed line).

where $X_{1:m}$ is the matrix of the first m columns of X, such that ΔY can be represented in the basis X as

$$\Delta Y = \varepsilon X D, \qquad D = O(1),\ D \in \mathbb{R}^{n\times m}.$$

then

$$\lambda_{j+1} - \lambda_j \ = \ O(1), \quad j = 1, \ldots, n-1. \tag{11}$$

This assumption is always valid if the SCF iteration is sufficiently converged such that ε is small.

The error of the new approximation

$$Y_{\mathrm{new}} \ = \ X_{1:m} + \Delta Y_{\mathrm{new}} \tag{12}$$

is analogously represented as

$$\Delta Y_{\mathrm{new}} \ = \ \varepsilon X D_{\mathrm{new}}. \tag{13}$$

Our goal is to derive bounds for $\|d_{\mathrm{new},j}\|$ in terms of $\|d_j\|$ where $d_{\mathrm{new},j}$ and d_j $(j = 1, \ldots, m)$ denote (the relevant parts[1] of) the columns of D_{new} and D, respectively. Then, bounds for the new error $\|\Delta y_{\mathrm{new},j}\|_{X^TS}$ in terms of the old error $\|\Delta y_j\|_{X^TS}$ in the norm $\|\cdot\|_{X^TS}$ defined by $\|u\|_{X^TS} = \|X^TSu\| = \|X^{-1}u\|$ follow immediately. Here and throughout the paper, $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices. Our main result is summarized by the following theorem, whose lengthy and technical proof is relegated to Appendix A.

Theorem 1 *The error of the new approximation to the j-th eigenvector can be estimated in terms of the error of the old approximation by*

$$\|\Delta y_{\mathrm{new},j}\|_{X^TS} \leq \begin{cases} \frac{|\lambda_j|}{\lambda_{m+1}}\|\Delta y_j\|_{X^TS} + O(\varepsilon^2) \text{ for } \lambda_j/\lambda_{m+1} \leq 2\sqrt{2}-2, \\ \frac{1}{4}\frac{(2-\lambda_j/\lambda_{m+1})^2}{\sqrt{1-\lambda_j/\lambda_{m+1}}}\|\Delta y_j\|_{X^TS} + O(\varepsilon^2) \text{ for } \lambda_j/\lambda_{m+1} \geq 2\sqrt{2}-2, \end{cases} \tag{14}$$

[1] Refer to the definitions of $k_j = d_{\mathrm{new},j}$ and d_j in Section A.2 of the Appendix.

$j = 1, \ldots, m$. *In the special case* $m = 1$ *it holds*

$$\|\Delta y_{\mathrm{new},1}\|_{X^T S} \leq \frac{|\lambda_1|}{\lambda_2} \|\Delta y_1\|_{X^T S} + O(\varepsilon^2) \tag{15}$$

even for $\lambda_1/\lambda_2 \geq 2\sqrt{2} - 2$.

By dividing the bounds from the theorem by $\|\Delta y_j\|_{X^T S}$ we obtain bounds (up to terms of order $O(\varepsilon)$) for the convergence rates $\|\Delta y_{\mathrm{new},j}\|_{X^T S}/\|\Delta y_j\|_{X^T S}$ which depend only on λ_j/λ_{m+1}. These bounds are plotted in Figure 1.

In Section 4.1 below we are going to demonstrate that these bounds are sharp, i. e. the bend in the curve (14) is indeed observed in some problems.

4 Numerical Examples

4.1 Example 1

We show how an artificial example can be contructed such that the observed convergence rates corresponding to *all* eigenvalues λ_j satisfying $-1 < \lambda_j/\lambda_{m+1} \leq 2\sqrt{2} - 2$, $j = 1, \ldots, m$ uniformly approach the bound (14) of Theorem 1. This illustrates the sharpness of this bound, and even its optimality in the following sense: For $-1 < \lambda_j/\lambda_{m+1} \leq 2\sqrt{2} - 2$ it is the best possible under all bounds only depending on λ_j and λ_{m+1}.

We choose $m = 100$ and $n = 3m$. We fix the first m eigenvalues $\lambda_1, \ldots, \lambda_m \in (-1, 1) \setminus \{0\}$ arbitrarily, the next m eigenvalues $\lambda_{m+1}, \ldots, \lambda_{2m} > 1$ very close to 1, and the last m eigenvalues $\lambda_{2m+1}, \ldots, \lambda_{3m}$ very large. Concretely, we choose $\lambda_1, \ldots, \lambda_m$ evenly spaced in $[-0.99, 0.99]$,

$$\lambda_j = -0.99 + \frac{2(j-1)}{m}, \quad j = 1, \ldots, m,$$

and the other eigenvalues as

$$\lambda_{m+j} = 1 + j\delta_1, \quad j = 1, \ldots, m \quad \text{with } \delta_1 = 10^{-5}$$

and

$$\lambda_{2m+j} = j\delta_2^{-1}, \quad j = 1, \ldots, m \quad \text{with } \delta_2 = 10^{-11}.$$

For the matrix X of exact eigenvectors we choose a perturbed unit matrix

$$X = \begin{bmatrix} I_m & 0 & 0 \\ 0 & \alpha I_m & \beta I_m \\ 0 & -\beta I_m & \alpha I_m \end{bmatrix} \quad \text{with } \delta_3 = 10^{-4}, \quad \alpha = \sqrt{1 - \delta_3^2}, \quad \beta = \delta_3,$$

which is orthogonal by construction. The data matrices H, S in (1) are then defined as

$$H = X \operatorname{diag}(\lambda_1, \ldots, \lambda_{3m}) X^T, \quad S = I_{3m}.$$

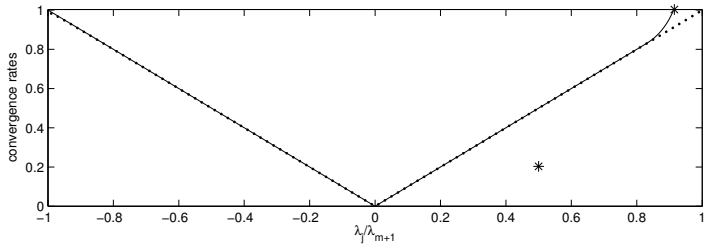

Fig. 2 Numerically observed convergence rates $\|\Delta y_{\mathrm{new},j}\|_{X^TS}/\|\Delta y_j\|_{X^TS}$ of Example 1 (dots) and of Example 2 (stars) Also shown is the bound for the convergence rates given by Theorem 1, eq. (14) (line).

For the initial approximations $Y = [y_1, \ldots, y_m]$ to the exact eigenvectors $X_{1:m} = [x_1, \ldots, x_m]$ we choose

$$Y = X_{1:m} + \varepsilon \begin{bmatrix} 0 \\ I_m \\ 0 \end{bmatrix} \quad \text{with } \varepsilon = 10^{-3},$$

where both zero sub-matrices are $\in \mathbb{R}^{m\times m}$, so that clearly

$$err_j = \|y_j - x_j\| = \varepsilon, \quad j = 1, \ldots, m.$$

The parameters $\delta_1, \delta_2, \delta_3, \varepsilon$ have to be carefully chosen (i.e., chosen not too small) to avoid numerical difficulties like cancellation or overflow. Furthermore, to make sure that the matrix H is not numerically singular the eigenvalues should not be chosen too close to zero, which for our evenly spaced eigenvalues would be the case if m were odd or too large. With our particular setting of the parameters everything works fine using MATLAB on standard hardware. Our algorithm computes new approximations $Y_{\mathrm{new}} = [y_{\mathrm{new},1}, \ldots, y_{\mathrm{new},m}]$ to the eigenvectors with errors

$$err_{\mathrm{new},j} = \|y_{\mathrm{new},j} - (\pm x_j)\|, \quad j = 1, \ldots m,$$

where the factor ± 1 selects the eigenvector such that $y_{\mathrm{new},j}$ shows the smaller error. The numerically observed convergence rates $err_{\mathrm{new},j}/err_j$, $j = 1, \ldots, m$ are plotted in Figure 2. Note that for this example the bound given by Theorem 1, eq. (14) is uniformly very sharp as long as $\lambda_j/\lambda_{m+1} \leq 2\sqrt{2}-2 \doteq 0.8284$.

4.2 Example 2

We construct an example where for the eigenvector corresponding to an eigenvalue λ_j with $-1 < \lambda_j/\lambda_{m+1} < 1$ our algorithm yields a convergence rate > 1. Thus, this example illustrates that in fact the error may increase for certain pathological problem data and hence our estimate in Theorem 1 is sharp also

in such a situation. In practical computations, see for example [2], such a behavior has never been encountered, however. We set $n = 5$, $m = 2$, i.e., the dimensions as small as possible (because for $m = 1$ Theorem 1, eq. (15) would guarantee a convergence rate < 1).

We choose

$$H = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) = \mathrm{diag}(0.5, 0.915, 1, 1.5, 10000), \quad S = I_5.$$

Note that $\lambda_2/\lambda_3 > 0.9126 > 2\sqrt{2} - 2 \doteq 0.8284$, where $x \doteq 0.9126$ is a solution of $\frac{1}{4}\frac{(2-x)^2}{\sqrt{1-x}} = 1$ such that the second branch in the piecewise-defined estimate (14) applies, and such that (14) does not exclude a convergence rate > 1, i. e., an increase in the approximation error. The exact eigenvectors are the canonical unit vectors,

$$X = [x_1, x_2, x_3, x_4, x_5] = I_5.$$

As the initial approximations to the eigenvectors x_1, x_2 we choose

$$Y = [y_1, y_2] = \begin{bmatrix} 1.000000000000000 & 0.000000000000000 \\ 0.000000000000000 & 1.000000000000000 \\ 0.000613604339291 & 0.000624080400796 \\ -0.000083591341207 & 0.000780017095933 \\ 0.000014803795114 & 0.000045792831252 \end{bmatrix},$$

such that the errors of these approximations are

$$err_1 = \|y_1 - x_1\| \approx 6.194 \cdot 10^{-4},$$
$$err_2 = \|y_2 - x_2\| \approx 1.0 \cdot 10^{-3},$$

respectively. For this data our algorithm computes new approximations

$$Y_{\mathrm{new}} = [y_{\mathrm{new},1}, y_{\mathrm{new},2}] \doteq \begin{bmatrix} 0.999999992092387 & -0.000000050401176 \\ -0.000000161788990 & -0.999999497314401 \\ 0.000091632309098 & -0.000967246231786 \\ 0.000086131966404 & -0.000264207603769 \\ -0.000000062534618 & 0.000000112221290 \end{bmatrix},$$

to the eigenvectors with errors

$$err_{\mathrm{new},1} = \|y_{\mathrm{new},1} - x_1\| \approx 1.258 \cdot 10^{-4},$$
$$err_{\mathrm{new},2} = \|y_{\mathrm{new},2} - (-x_2)\| \approx 1.00268 \cdot 10^{-3}.$$

Note that here $y_{\mathrm{new},2}$ is an approximation to $-x_2$, but the (real) normed eigenvectors of H are defined only up to a factor ± 1 anyway. For the convergence rates we obtain

$$\frac{err_{\mathrm{new},1}}{err_1} \approx 0.203 \leq \frac{\lambda_1}{\lambda_3} = 0.5,$$
$$\frac{err_{\mathrm{new},2}}{err_2} \approx 1.00268 \leq \frac{1}{4}\frac{(2 - \lambda_2/\lambda_3)^2}{\sqrt{1 - \lambda_2/\lambda_3}} \approx 1.00946,$$

i.e., for the second eigenvector a convergence rate > 1 with a very sharp bound (14). These convergence rates are plotted in Figure 2.

4.3 Example 3

We consider a *nonlinear* eigenvalue problem of the form

$$H(X)X = X\Lambda_m, \tag{16}$$

where $X \in \mathbb{R}^{n\times m}$, $X^TX = I_m$, $H(X) \in \mathbb{R}^{n\times n}$ is a symmetric matrix depending on X, and $\Lambda_m \in \mathbb{R}^{m\times m}$ is a diagonal matrix containing the m smallest eigenvalues of $H(X)$. The (discretized) Hartree-Fock and Kohn-Sham equations in electronic structure calculations are essentially of this type. Following [28] we consider a simplified version of these equations[2]. The dependency of $H(X)$ on X is expressed through the vector

$$\rho(X) = \mathrm{diag}(XX^T) \in \mathbb{R}^n \tag{17}$$

containing the diagonal elements of the matrix XX^T, which in electronic structure calculations would correspond to the charge density of electrons. Then $H(X)$ is defined by

$$H(X) = L + \alpha\,\mathrm{diag}(L^{-1}\rho(X)), \tag{18}$$

where $\alpha \in \mathbb{R}$ and $L \in \mathbb{R}^{n\times n}$ denotes a discrete version of the Laplace operator.

For the numerical solution of the nonlinear eigenvalue problem (16)–(18) we apply a version of the self-consistent field (SCF) iteration, which is given by the following algorithm:

Input: An initial guess $X^{(0)}$ satisfying $(X^{(0)})^TX^{(0)} = I_m$ for the solution X of (16);
Output: Numerical solution X of (16);
for $l = 1, 2, \ldots$ until convergence **do**
  Construct $H^{(l)} = H(X^{(l-1)})$ according to (17), (18);
  Solve the *linear* eigenproblem $H^{(l)}X^{(l)} = X^{(l)}\Lambda_m^{(l)}$ such that $(X^{(l)})^TX^{(l)} = I_m$ and $\Lambda_m^{(l)}$ contains the m smallest eigenvalues of $H^{(l)}$;
end do
Set $X =$ the last computed $X^{(l)}$;

We refer to this procedure as the "SCF iteration with full diagonalization" to distinguish it from the following variant referred to as "SCF iteration with approximate diagonalization (3)", where the respective linear eigenproblems

$$H^{(l)}X^{(l)} = X^{(l)}\Lambda_m^{(l)} \tag{19}$$

are solved only approximately by employing our algorithm described in Section 2. Here, the approximation $X^{(l-1)}$ computed in the previous iteration step is used for the old approximation to the eigenvectors required by this

[2] Our method's performance for realistic applications in electronic structure computations is thoroughly discussed in [2].

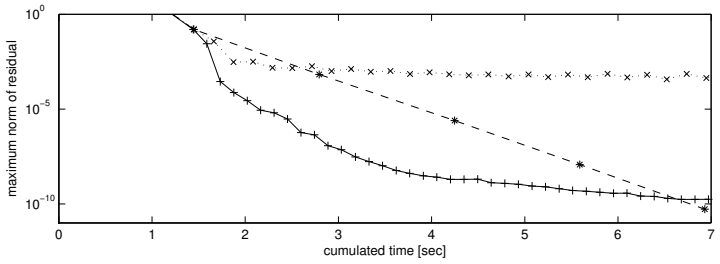

Fig. 3 Residual versus cumulative computing time for the SCF iteration with full (∗) and approximate (+) diagonalization; Davidson variant (×).

algorithm. However, in the first iteration step ($l = 1$) a reasonably good approximation $X^{(0)}$ is usually not available. Consequently, for $l = 1$ we compute the full solution $X^{(1)}$ of (19) and apply the approximate procedure only for $l \geq 2$.

As a third algorithm in this comparison we use the Davidson method [6] which was originally implemented in the WIEN2k code, see (8).

Figure 3 illustrates a typical progress of the SCF iteration with both full and approximate diagonalization. Here we choose $n = 1000$, $m = 30$, $\alpha = 0.1$, and for L the discrete 1D-Laplace operator on $[0, 10]$ with Dirichlet boundary conditions, i.e.,

$$L = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n \times n}, \quad \text{where } h = \frac{10}{n}.$$

For the initial guess we choose $X^{(0)} = [I_m \, 0]^T$ with the zero matrix $0 \in \mathbb{R}^{m \times (n-m)}$. The figure shows the maximum norm (the maximum of absolute values over all entries) of the residual $H(X^{(l)})X^{(l)} - X^{(l)}\Lambda_m^{(l)}$ versus the cumulative computing time for the first 5 (full diagonalization) respectively 39 (approximate diagonalization) SCF iteration steps. Here only the qualitative picture is of relevance, not the particular timing of our MATLAB implementation on current standard hardware. We observe that in this example the improvement per iteration step is significantly better for the SCF iteration with full diagonalization, but, more importantly, as long as only moderate precision is required, the total improvement until a certain cumulated computing time is significantly better for the variant with our new approximate diagonalization, although more iteration steps are possibly necessary. This is due to the fact that the cost of one SCF iteration step with approximate diagonalization is dominated by the Cholesky decomposition of $H^{(l)}$ needed for (3), which is much cheaper than a full solution of (19), which dominates the cost of one SCF iteration step with full diagonalization. Moreover, we observe

that the classical Davidson method does not provide a sufficiently accurate approximation even in this simple example.

5 Conclusions

In this paper, we have analyzed a subspace method applicable for the approximate solution of generalized eigenvalue problems in linear algebra as they arise for instance in large–scale DFT computations of electronic structure, where the subspace is expanded based on a new preconditioner. We have derived estimates of the improvement achieved by our method starting from a suitable initial approximation. Numerical examples show that the estimates we derived are sharp and also apply in realistic examples from applications.

Acknowledgements

We would like to thank P. Blaha, R. Laskowski and K. Schwarz from the Theoretical Chemistry group of the Institute for Materials Science at Vienna University of Technology for valuable discussions on the method analyzed in this paper and the possibility to realize and test the algorithm in the WIEN2k code [2].

A Proof of Theorem 1

In this Appendix we give the proof of Theorem 1. First, in Subsection A.1 we characterize D_{new} from (13) as the solution of a certain Sylvester equation. Then, in Subsection A.2 we derive an explicit representation of the solution of this equation. The proof of Theorem 1 is thus reduced to the derivation of bounds for this solution, which we state as Proposition 3 and prove for the special case $m = 1$ in Subsection A.3 and for the special case $n = 2m+1$ in Subsection A.4. Finally, in Subsection A.5 we show that the general case can be reduced to the latter special one, which completes the proof of Proposition 3 and thus also of Theorem 1.

A.1 Characterization of D_{new} as the Solution of a Sylvester Equation

Proposition 1 *Let*

$$D = \begin{bmatrix} D_1 \\ D_2 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}, \qquad \hat{D}_2 := \Lambda_2^{-1} D_2 \Lambda_1 \tag{20}$$

with $D_1 \in \mathbb{R}^{m\times m}$, $D_2 \in \mathbb{R}^{(n-m)\times m}$, $\Lambda_1 \in \mathbb{R}^{m\times m}$, $\Lambda_2 \in \mathbb{R}^{(n-m)\times(n-m)}$. Let P denote the orthogonal projection of $\mathbb{R}^{n-m}$ onto the subspace spanned by the columns of $D_2 - \hat{D}_2$, i.e.,

$$P^T = P, \quad P^2 = P, \quad P(D_2 - \hat{D}_2) = D_2 - \hat{D}_2.$$

Then for D_{new} given in (13)*,*

$$D_{\text{new}} = \begin{bmatrix} 0 \\ K \end{bmatrix} + O(\varepsilon), \tag{21}$$

holds, where $0 \in \mathbb{R}^{m\times m}$ and $K \in \mathbb{R}^{(n-m)\times m}$ is the unique solution of the Sylvester equation

$$P\Lambda_2 K - K\Lambda_1 + (I_{n-m} - P)D_2\Lambda_1 = 0. \tag{22}$$

The proof of the proposition is constituted by the following consecutive substeps.

Step 1 *The Ritz values* (2) *satisfy*

$$\vartheta_j = \lambda_j + O(\varepsilon^2), \qquad j = 1, \ldots, m.$$

Proof

$$\begin{aligned}
\vartheta_j = \frac{y_j^T H y_j}{y_j^T S y_j} &= \frac{(x_j + \Delta y_j)^T(\lambda_j S x_j + H\Delta y_j)}{(x_j + \Delta y_j)^T(S x_j + S\Delta y_j)} \\
&= \frac{\lambda_j x_j^T S x_j + \lambda_j \Delta y_j^T S x_j + x_j^T H \Delta y_j + O(\varepsilon^2)}{x_j^T S x_j + \Delta y_j^T S x_j + x_j^T S \Delta y_j + O(\varepsilon^2)} \\
&= \frac{\lambda_j(1 + 2\alpha_j) + O(\varepsilon^2)}{1 + 2\alpha_j + O(\varepsilon^2)} = \lambda_j + O(\varepsilon^2).
\end{aligned}$$

Here we used $\alpha_j := x_j^T S \Delta y_j$ and $x_j^T H \Delta y_j = \lambda_j x_j^T S \Delta y_j$, and noted (9). □

Step 2 *The representation of* Z *from* (3) *in the basis* X *satisfies*

$$Z = \varepsilon X(D - \hat{D}) + O(\varepsilon^2) \quad \textit{with} \quad \hat{D} := \Lambda^{-1} D \Lambda_1. \tag{23}$$

Proof

$$\begin{aligned}
z_j &= H^{-1}(H - \vartheta_j S) y_j \\
&= y_j - \vartheta_j H^{-1} S y_j \\
&= x_j + \Delta y_j - \lambda_j (H^{-1} S x_j + H^{-1} S \Delta y_j) + O(\varepsilon^2) \\
&= x_j + \Delta y_j - \lambda_j (\frac{1}{\lambda_j} x_j + H^{-1} S \Delta y_j) + O(\varepsilon^2) \\
&= \Delta y_j - \lambda_j H^{-1} S \Delta y_j + O(\varepsilon^2) \\
&= \varepsilon (X d_j - \lambda_j H^{-1} S X d_j) + O(\varepsilon^2) \\
&= \varepsilon (X d_j - \lambda_j X \Lambda^{-1} d_j) + O(\varepsilon^2).
\end{aligned}$$

Here, d_j are the columns of D. □

Step 3 *Let*

$$U := X^{-1}[Y \;\; Z].$$

Then the matrices $\tilde{H}$ *and* $\tilde{S}$ *from* (5) *and* (6) *can be written as*

$$\tilde{H} = U^T \Lambda U \quad \textit{and} \quad \tilde{S} = U^T U \tag{24}$$

and it holds

$$U = [I_{n,m} + \varepsilon D, \quad \varepsilon(D - \hat{D})] + O(\varepsilon^2), \qquad \hat{D} := \Lambda^{-1} D \Lambda_1. \tag{25}$$

Here $I_{n,m} \in \mathbb{R}^{n \times m}$ *consists of the first* m *columns of* I_n.

Proof From $X^T H X = X^T S X \Lambda = \Lambda$ (cf. (1) and (9)) it follows

$$\tilde{H} = [Y \;\; Z]^T H [Y \;\; Z] = U^T X^T H X U = U^T \Lambda U$$

and

$$\tilde{S} = [Y \;\; Z]^T S [Y \;\; Z] = U^T X^T S X U = U^T U.$$

(25) is a consequence of $Y = X(I_{n,m} + \varepsilon D)$ and (23). □

Next, we solve the reduced eigenproblem (4) by transforming it to an eigenproblem in standard form. We achieve this by a suitable factorization of the matrix U in (24). Note that $U = X^{-1}[Y\ Z]$ has full rank $2m$ by assumption (cf. the end of Section 2).

Step 4 *Let the columns of $Q \in \mathbb{R}^{n\times 2m}$ form an orthonormal basis of the subspace spanned by the columns of U, i.e.,*

$$Q^T Q = I_{2m} \quad \textit{and} \quad U = Q\Phi \tag{26}$$

for an invertible coordinate transformation matrix $\Phi \in \mathbb{R}^{2m\times 2m}$. Further let $\Omega_{1:m} \in \mathbb{R}^{2m\times m}$ consist of orthonormal eigenvectors corresponding to the m lowest eigenvalues $\gamma_1 \le \cdots \le \gamma_m$ of the symmetric eigenproblem (in standard form)

$$Q^T \Lambda Q \Omega = \Omega \Gamma. \tag{27}$$

Then there is a solution $V_{1:m}$ of (4) such that Y_{new} defined by (7) satisfies

$$Y_{\text{new}} = XQ\Omega_{1:m}. \tag{28}$$

Proof Let

$$V_{1:m} := \Phi^{-1}\Omega_{1:m}.$$

Then the columns of $V_{1:m}$ are orthonormal with respect to $\tilde{S}$,

$$\begin{aligned} V_{1:m}^T \tilde{S} V_{1:m} &= \Omega_{1:m}^T \Phi^{-T} \tilde{S} \Phi^{-1} \Omega_{1:m} \\ &= \Omega_{1:m}^T \Phi^{-T} U^T U \Phi^{-1} \Omega_{1:m} \\ &= \Omega_{1:m}^T Q^T Q \Omega_{1:m} \ = \ \Omega_{1:m}^T \Omega_{1:m} \ = \ I_m, \end{aligned}$$

and it holds

$$\begin{aligned} \tilde{H} V_{1:m} &= U^T \Lambda U \Phi^{-1} \Omega_{1:m} = \Phi^T Q^T \Lambda Q \Omega_{1:m} = \Phi^T \Omega_{1:m} \Gamma \\ &= \Phi^T \Phi V_{1:m} \Gamma = \Phi^T Q^T Q \Phi V_{1:m} \Gamma = U^T U V_{1:m} \Gamma \\ &= \tilde{S} V_{1:m} \Gamma. \end{aligned}$$

So $V_{1:m}$ is a solution of (4) and it holds

$$Y_{\text{new}} \ = \ [Y\ Z] V_{1:m} \ = \ XU\Phi^{-1} W_{1:m} \ = \ XQW_{1:m}.$$

□

Let us stress that in (26) $U = Q\Phi$ holds exactly (and not just up to terms of order $O(\varepsilon^2)$). In the previous proof all calculations are exact and thus (28) holds exactly, too. Note that the fact that the entries of the matrix Φ^{-1} are possibly unbounded for $\varepsilon \to 0$ does not affect the arguments.

An orthonormal basis Q of U can be obtained by computing a QR-decomposition of U. However, to obtain an orthonormal basis with more favorable properties for the further analysis, we first perform some elementary column transformations on U, then compute the QR-decomposition $UJ = QR$ of the transformed matrix UJ, and finally set $\Phi := RJ^{-1}$ in (26).

Step 5 *There exists $J \in \mathbb{R}^{2m\times 2m}$ with $J = O(1)$ and $J^{-1} = O(1)$ such that*

$$UJ = \begin{bmatrix} I_m & 0 \\ \varepsilon D_2 & \varepsilon(D_2 - \hat{D}_2) \end{bmatrix} + O(\varepsilon^2). \tag{29}$$

Here D_2 and $\hat{D}_2 = \Lambda_2^{-1} D_2 \Lambda_1$ are defined as in (20).

Proof Using (25) it is easily verified that (29) holds for

$$J := \begin{bmatrix} I_m - \varepsilon D_1 & -\varepsilon(D_1 - \hat{D}_1) \\ 0 & I_m \end{bmatrix}$$

and, consequently,

$$J^{-1} = \begin{bmatrix} I_m + \varepsilon D_1 & \varepsilon(D_1 - \hat{D}_1) \\ 0 & I_m \end{bmatrix} + O(\varepsilon^2),$$

where $\hat{D}_1 = \Lambda_1^{-1} D_1 \Lambda_1$. □

Step 6 *For UJ a QR-decomposition*

$$UJ = QR \tag{30}$$

(where $Q \in \mathbb{R}^{n\times 2m}$ satisfies $Q^TQ = I_{2m}$ and $R \in \mathbb{R}^{2m\times 2m}$ is an upper triangular matrix) can be chosen such that Q is of the form

$$Q = \begin{bmatrix} I_m + O(\varepsilon^2) & \varepsilon Q_{12} + O(\varepsilon^2) \\ \varepsilon Q_{21} + O(\varepsilon^2) & Q_{22} + O(\varepsilon) \end{bmatrix}. \tag{31}$$

Here $Q_{12} \in \mathbb{R}^{m\times m}$, $Q_{21} \in \mathbb{R}^{(n-m)\times m}$, and $Q_{22} \in \mathbb{R}^{(n-m)\times m}$ are of order $O(1)$ and satisfy

$$Q_{21} = D_2, \tag{32}$$
$$Q_{12} = -Q_{21}^T Q_{22}, \tag{33}$$
$$Q_{22}^T Q_{22} = I_m, \tag{34}$$
$$Q_{22} Q_{22}^T = P, \tag{35}$$

where P is the orthogonal projection onto $D_2 - \hat{D}_2$.

Proof Q is of the form

$$Q = \begin{bmatrix} Q_{11}^0 + \varepsilon Q_{11}^1 + O(\varepsilon^2) & Q_{12}^0 + \varepsilon Q_{12} + O(\varepsilon^2) \\ Q_{21}^0 + \varepsilon Q_{21} + O(\varepsilon^2) & Q_{22} + O(\varepsilon) \end{bmatrix}.$$

We want to show that $Q_{12}^0 = 0$, $Q_{21}^0 = 0$, $Q_{11}^1 = 0$, and $Q_{11}^0 = I_m$. Let

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}.$$

By comparing the lower left blocks in

$$QR = UJ = \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} + O(\varepsilon)$$

(cf. (29)) up to terms of order $O(1)$, we obtain $Q_{21}^0 R_{11} = 0$, so $Q_{21}^0 = 0$, because R_{11} is non-singular. Then the upper left blocks in $Q^TQ = I_{2m}$ and $QR = UJ$ give respectively

$${Q_{11}^0}^T Q_{11}^0 = I_m \tag{36}$$

and $I_m = (Q_{11}^0 + \varepsilon Q_{11}^1)R_{11} + O(\varepsilon^2)$, thus $Q_{11}^1 = 0$ and

$$Q_{11}^0 R_{11} = I_m.$$

R_{11} is an upper triangular matrix, so $Q_{11}^0 = R_{11}^{-1}$ is also upper triangular. On the other hand, $(Q_{11}^0)^T = (Q_{11}^0)^{-1} = R_{11}$, so Q_{11}^0 is both upper and lower triangular, i.e., Q_{11}^0 is a diagonal matrix with ± 1 elements as follows from (36). So for some choice of the QR

decomposition (30) it holds $Q_{11}^0 = I_m$. Finally, $Q_{12}^0 = 0$ now follows from comparing the right upper blocks in $Q^T Q = I_{2m}$.

The identities (33) and (34) follow from (31) by evaluating $Q^T Q = I_{2m}$ up to terms of order $O(\varepsilon)$.

Now, by comparing the upper left blocks in $QR = UJ$ up to terms of order $O(1)$ we obtain $(I_m + O(\varepsilon^2))R_{11} = I + O(\varepsilon)$ and thus $R_{11} = I + O(\varepsilon)$. Substituting this into $\varepsilon Q_{21} R_{11} = \varepsilon D_2 + O(\varepsilon^2)$ (the lower left block of $QR = UJ$ up to terms of order $O(\varepsilon)$) we obtain (32).

Similarly, comparing the upper right blocks in $QR = UJ$ up to terms of order $O(1)$ yields $(I_m + O(\varepsilon^2))R_{12} = O(\varepsilon)$ and thus $R_{12} = O(\varepsilon)$. Substituting this and $Q_{21} = D_2$ into $\varepsilon Q_{21} R_{12} + Q_{22} R_{22} = \varepsilon(D_2 - \hat{D}_2) + O(\varepsilon^2)$ (the lower right block of $QR = UJ$) we see from (34) that $Q_{22} R_{22}$ can be chosen as a QR-decomposition of $\varepsilon(D_2 - \hat{D}_2)$. From this it follows that $P := Q_{22} Q_{22}^T$ is the orthogonal projection onto $D_2 - \hat{D}_2$. □

Step 7 *The eigenvector matrix $\Omega \in \mathbb{R}^{2m \times 2m}$ in* (27) *with Q defined in* (30) *can be chosen in the form*

$$\Omega \;=\; \begin{bmatrix} I_m + O(\varepsilon^2) & \varepsilon\Omega_{12} + O(\varepsilon^2) \\ \varepsilon\Omega_{21} + O(\varepsilon^2) & \Omega_{22} + O(\varepsilon) \end{bmatrix}, \tag{37}$$

where $\Omega_{12} \in \mathbb{R}^{m \times m}$, $\Omega_{21} \in \mathbb{R}^{m \times m}$, and $\Omega_{22} \in \mathbb{R}^{m \times m}$ are of order $O(1)$ and satisfy

$$\Omega_{12} = -\Omega_{21}^T \Omega_{22}, \tag{38}$$

$$\Omega_{21} = -\Omega_{22} \Omega_{12}^T, \tag{39}$$

$$\Omega_{22}^T \Omega_{22} \;=\; \Omega_{22} \Omega_{22}^T = I_m. \tag{40}$$

Proof From (31) it follows

$$Q^T \Lambda Q = \begin{bmatrix} \Lambda_1 + O(\varepsilon) & O(\varepsilon) \\ O(\varepsilon) & Q_{22}^T \Lambda_2 Q_{22} + O(\varepsilon) \end{bmatrix}.$$

Since the eigenvectors of $\begin{bmatrix} \Lambda_1 & 0 \\ 0 & Q_{22}^T \Lambda_2 Q_{22} \end{bmatrix}$ have the form $\begin{bmatrix} I_m & 0 \\ 0 & \Omega_{22} \end{bmatrix}$, it follows from the well-conditioning of the symmetric eigenproblem under the assumption (11) that the matrix Ω consisting of the eigenvectors of $Q^T \Lambda Q$ is of the form $\begin{bmatrix} I_m + O(\varepsilon) & O(\varepsilon) \\ O(\varepsilon) & \Omega_{22} + O(\varepsilon) \end{bmatrix}$, i.e.,

$$\Omega \;=\; \begin{bmatrix} I_m + \varepsilon G + O(\varepsilon^2) & \varepsilon\Omega_{12} + O(\varepsilon^2) \\ \varepsilon\Omega_{21} + O(\varepsilon^2) & \Omega_{22} + O(\varepsilon) \end{bmatrix}.$$

Here G is antisymmetric, $G^T = -G$, and equations (38), (39), and (40) hold, which follows from the evaluation of $\Omega^T \Omega = I_{2m}$ and $\Omega \Omega^T = I_{2m}$ up to terms of order $O(\varepsilon)$. We now show that actually $G = 0$. With

$$Q\Omega \;=\; \begin{bmatrix} I_m + \varepsilon G + O(\varepsilon^2) & \varepsilon\Omega_{12} + \varepsilon Q_{12}\Omega_{22} + O(\varepsilon^2) \\ \varepsilon Q_{21} + \varepsilon Q_{22}\Omega_{21} + O(\varepsilon^2) & Q_{22}\Omega_{22} + O(\varepsilon) \end{bmatrix} \tag{41}$$

the upper left block in $\Omega^T Q^T \Lambda Q \Omega = \Gamma$ evaluated up to terms of order $O(\varepsilon)$ gives

$$\Lambda_1 + \varepsilon G^T \Lambda_1 + \varepsilon G \Lambda_1 + O(\varepsilon^2) \;=\; \Gamma_1,$$

and the antisymmetry of G shows that G and Λ_1 commute,

$$G\Lambda_1 \;=\; \Lambda_1 G.$$

From this $G = 0$ follows by an application of the following Lemma 1. □

Lemma 1 *Let $A \in \mathbb{R}^{n\times n}$ be antisymmetric,*

$$A^T \;=\; -A,$$

and let $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with distinct λ_j. If A and Λ commute,

$$A\Lambda \;=\; \Lambda A, \tag{42}$$

then

$$A \;=\; 0. \tag{43}$$

The lemma is proven by induction on n. □

Finally, we can give the proof of Proposition 1.

Step 8 (Proof of Proposition 1)**.** Let

$$K \;:=\; Q_{21} + Q_{22}W_{21} \;= D_2 + Q_{22}W_{21}.$$

Then

$$D_{\mathrm{new}} \;=\; \begin{bmatrix} 0 \\ K \end{bmatrix} + O(\varepsilon)$$

follows from (12), (13), (28), and (41). We have to show that K solves the Sylvester equation

$$P\Lambda_2 K - K\Lambda_1 + (I_{n-m} - P)D_2\Lambda_1 \;=\; 0.$$

Comparison of the upper right blocks in $W^TQ^T\Lambda QW = \Gamma$ up to terms of order $O(\varepsilon)$ yields

$$(W_{12}^T + W_{22}^TQ_{12}^T)\Lambda_1 + W_{22}^TQ_{22}^T\Lambda_2 K \;=\; 0.$$

With $W_{12}^T = -W_{22}^TW_{21}$ and $Q_{12}^T = -Q_{22}^TQ_{21}$ (cf. Steps 6 and 7) it follows

$$(-W_{22}^TW_{21} - W_{22}^TQ_{22}^TQ_{21})\Lambda_1 + W_{22}^TQ_{22}^T\Lambda_2 K \;=\; 0.$$

Multiplication with $Q_{22}W_{22}$ from the left using $W_{22}W_{22}^T = I_m$, $Q_{22}Q_{22}^T = P$, and $Q_{21} = D_2$ gives

$$(-Q_{22}W_{21} - PD_2)\Lambda_1 + P\Lambda_2 K \;=\; 0$$

or

$$\underbrace{(-Q_{22}W_{21} - D_2)}_{=-K}\Lambda_1 + P\Lambda_2 K \;=\; (PD_2 - D_2)\Lambda_1.$$

It remains to prove uniqueness of the solution K. Generally, a Sylvester equation

$$AX + XB + C \;=\; 0$$

has a unique solution, if and only if

$$\alpha + \beta \;\neq\; 0$$

for all eigenvalues α of A and all eigenvalues β of B. Now observe that the eigenvalues of $-\Lambda_1$ are $-\lambda_1, \ldots, -\lambda_m$ and all non-zero eigenvalues of $P\Lambda_2$ are $\geq \lambda_{m+1}$, see Lemma 2 below. Thus the uniqueness of K follows from the assumption that the λ_j are non-zero and distinct. □

Lemma 2 *Let $P \in \mathbb{R}^{n\times n}$ be an orthogonal projection, i.e.,*

$$P^T \;=\; P, \quad P^2 \;=\; P,$$

and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_j \in \mathbb{R}$ and $\lambda_j > 0$ for all $j = 1, \ldots, n$. Then all non-zero eigenvalues of $P\Lambda$ are real and lie in the interval $[\min_j \lambda_j, \max_j \lambda_j]$.

Proof If $P = I_n$ or $P = 0$ then the statement of the lemma clearly holds. So let $P \neq I_n$ and $P \neq 0$. Let γ be a non-zero eigenvalue of $P\Lambda$ with corresponding eigenvector x. Then

$$P\Lambda x \;=\; \gamma P x \quad \text{and} \quad Px \;=\; \frac{1}{\gamma} P\Lambda x \;=\; x,$$

and thus

$$P^T\Lambda P x \;=\; P\Lambda P x \;=\gamma x,$$

i.e., γ is an eigenvalue of the symmetric matrix $P^T\Lambda P$ and it follows $\gamma \in \mathbb{R}$. Moreover $\gamma > 0$ since $P^T\Lambda P$ is positive semidefinite and γ was assumed to be non-zero. It remains to prove that all non-zero eigenvalues of $P^T\Lambda P$ are in the interval $[\min_j \lambda_j, \max_j \lambda_j]$.

Since $P \neq 0, I_n$, P is a projection onto a proper subspace of $\mathbb{R}^n$ and thus P and $P^T\Lambda P$ are singular. It follows that 0 is the smallest eigenvalue of $P^T\Lambda P$. Let $\mathcal{N}$ be the null-space of P, i.e, the eigenspace corresponding to the smallest eigenvalue 0. Let γ_2 be the smallest non-zero eigenvalue of $P^T\Lambda P$. Then

$$\gamma_2 \;=\; \max_{y\in\mathcal{N}} \min_{x\neq 0,\; y^T x=0} \frac{x^T P^T \Lambda P x}{x^T x}$$

by the min-max principle for eigenvalues of symmetric matrices. Since $\mathcal{N} \perp P(\mathbb{R}^n)$, it holds $y^T P x = 0$ for all x and all $y \in \mathcal{N}$, and thus

$$\gamma_2 \geq \min_{x\neq 0,\; x=Px} \frac{x^T P^T \Lambda P x}{x^T x} = \min_{x\neq 0,\; x=Px} \frac{x^T \Lambda x}{x^T x} \geq \min_{x\neq 0} \frac{x^T \Lambda x}{x^T x} = \min_{j=1,\ldots,n} \lambda_j.$$

Finally, for the largest eigenvalue $\gamma_{\max}$ of $P^T\Lambda P$ we obtain

$$\gamma_{\max} = \max_{x\neq 0,\; x=Px} \frac{x^T \Lambda x}{x^T x} \leq \max_{x\neq 0} \frac{x^T \Lambda x}{x^T x} = \max_{j=1,\ldots,n} \lambda_j.$$

□

Remark. From (13) and (21) we obtain

$$\Delta Y_{\text{new}} \;=\; \varepsilon X_{m+1:n} K + O(\varepsilon^2). \tag{44}$$

Here it is quite remarkable that up to terms of order $O(\varepsilon^2)$ the new error ΔY_{new} has no components in the subspace spanned by the first m eigenvectors $X_{1:m}$. It is another notable fact, that up to terms of order $O(\varepsilon^2)$ the new error does not depend on components of the old error in the $X_{1:m}$-subspace, since (22) depends only on D_2 and not on D_1. Consequently, it is to be expected that in the first step the error will be reduced significantly if the algorithm is applied in an iterative way, since the components in the space $X_{1:m}$ are annihilated by the projection. Subsequent iteration steps will essentially operate in the $X_{m+1:n}$-subspace only, and the evolution of the error is described by the nonlinear operator $D_2 \mapsto K$ given in Proposition 1.

A.2 Explicit Representation of D_{new}

Next, we give an explicit representation of the solution K of the Sylvester equation (22) which by (21) constitutes an explicit representation of D_{new} up to terms of order $O(\varepsilon)$. Let d_j and k_j $(j = 1, \ldots, m)$ denote the j-th columns of D_2 and K, respectively. Similarly, let $\hat{d}_j = \lambda_j \Lambda_2^{-1} d_j$ denote the j-th column of $\hat{D}_2$.

Proposition 2 *Let $W \in \mathbb{R}^{(n-m)\times m}$ consist of orthonormal eigenvectors corresponding to the non-zero eigenvalues $\gamma_{m+1} \leq \cdots \leq \gamma_{2m}$ of the symmetric matrix $P\Lambda_2 P$,*

$$P\Lambda_2 P W \;=\; W\Gamma_2, \qquad \Gamma_2 = \operatorname{diag}(\gamma_{m+1}, \ldots, \gamma_{2m}), \tag{45}$$

where P denotes the projection onto $D_2 - \hat{D}_2$ as in Proposition 1. Then

$$k_j = \hat{d}_j - \lambda_j W(\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j), \quad j = 1, \ldots, m. \tag{46}$$

Proof We have to verify that

$$P\Lambda_2 k_j - \lambda_j k_j + \lambda_j (I_{n-m} - P) d_j = 0, \quad j = 1, \ldots, m,$$

cf. (22). This is done by a straightforward calculation using $\hat{d}_j = \lambda_j \Lambda_2^{-1} d_j$, $P = WW^T$, $W^T \Lambda_2 W = \Gamma_2$, and $P(d_j - \hat{d}_j) = d_j - \hat{d}_j$:

$$\begin{aligned}
P\Lambda_2 k_j - \lambda_j k_j &= (P\Lambda_2 - \lambda_j I_{n-m}) \left(\hat{d}_j - \lambda_j W (\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j) \right) \\
&= \lambda_j P d_j - \lambda_j \hat{d}_j - \lambda_j W W^T \Lambda_2 W (\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j) \\
&\qquad + \lambda_j^2 W (\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j) \\
&= \lambda_j P d_j - \lambda_j \hat{d}_j - \lambda_j W (\Gamma_2 - \lambda_j I_m)(\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j) \\
&= \lambda_j P d_j - \lambda_j \hat{d}_j - \lambda_j P (d_j - \hat{d}_j) \\
&= \lambda_j P d_j - \lambda_j \hat{d}_j - \lambda_j (d_j - \hat{d}_j) \\
&= \lambda_j (P - I_{n-m}) d_j .
\end{aligned}$$

□

Note that up to terms of order $O(\varepsilon)$ the j-th column of D_{new} has the form $\begin{bmatrix} 0 \\ k_j \end{bmatrix}$ and depends only on d_j, i.e., the first m entries of the j-th column $\begin{bmatrix} * \\ d_j \end{bmatrix}$ of D are irrelevant. It follows that the bounds of Theorem 1 are equivalent to corresponding bounds of $\|k_j\|$ in terms of $\|d_j\|$. Thus the proof of Theorem 1 is reduced to establishing the following Proposition 3.

Proposition 3 *For k_j from* (46)*, the following estimates hold:*

$$\|k_j\| \leq \begin{cases} \dfrac{|\lambda_j|}{\lambda_{m+1}} \|d_j\| & \text{for } \lambda_j/\lambda_{m+1} \leq 2\sqrt{2} - 2, \\ \dfrac{1}{4} \dfrac{(2 - \lambda_j/\lambda_{m+1})^2}{\sqrt{1 - \lambda_j/\lambda_{m+1}}} \|d_j\| & \text{for } \lambda_j/\lambda_{m+1} \geq 2\sqrt{2} - 2, \end{cases} \qquad j = 1, \ldots, m. \tag{47}$$

In the special case $m = 1$ it holds

$$\|k_1\| \leq \frac{|\lambda_1|}{\lambda_2} \|d_1\| \tag{48}$$

even for $\lambda_1/\lambda_2 \geq 2\sqrt{2} - 2$.

In the following Subsection A.3 we give the proof of (48) for the special case $m = 1$. Then, in Subsection A.4 we prove (47) in the special case $n = 2m + 1$, and finally, in Subsection A.5 we show how the general case can be reduced to this special one, and thus that (47) holds in general.

A.3 The Special Case $m = 1$

For $m = 1$, D_2 and $\hat{D}_2$ defined in Proposition 1 both consist of one column only which we denote by d and $\hat{d}$, respectively. Then also W defined in Proposition 2 consists of one column w only, which is a unit vector in the one-dimensional subspace spanned by $d - \hat{d}$, such that (up to sign) it must hold

$$w = \frac{d - \hat{d}}{\|d - \hat{d}\|}$$

and

$$\Gamma_2 \;=\; \gamma_2 \;=\; w^T \Lambda_2 w \;=\; \frac{\langle d-\hat{d}, \Lambda_2(d-\hat{d})\rangle}{\|d-\hat{d}\|^2}. \tag{49}$$

Now (46) of Proposition 2 reduces to

$$k \;=\; k_1 \;=\; \hat{d} - \frac{\lambda_1}{\gamma_2-\lambda_1}(d-\hat{d}) \;=\; \frac{-\lambda_1}{\gamma_2-\lambda_1} d + \frac{\gamma_2}{\gamma_2-\lambda_1}\hat{d}. \tag{50}$$

We will show that $\|k\| \le \|\hat{d}\|$ by applying the following simple lemma.

Lemma 3 *If $y \in \mathbb{R}^n$ lies on the line through two points p_1 and p_2 with $\|p_2\| < \|p_1\|$,*

$$y = (1-t)p_1 + tp_2 \quad \textit{for some } t \in \mathbb{R},$$

then

$$\begin{aligned} &\|y\| \le \|p_1\| \quad \textit{if and only if} \quad 0 \le t \le 2t_{\min}, \\ &\|y\| \le \|p_2\| \quad \textit{if and only if} \quad 1 \le t \le 2t_{\min} - 1. \end{aligned}$$

Here

$$t_{\min} \;=\; \frac{\langle p_1, p_1 - p_2\rangle}{\|p_1 - p_2\|^2}$$

is such that

$$y_{\min} = (1-t_{\min})p_1 + t_{\min} p_2$$

is the point on the line through p_1 and p_2 with minimal norm.

Proof Elementary analytical geometry. □

Proposition 3 (for the special case $m = 1$) *In the case $m = 1$ it holds*

$$\|k\| \le \|\hat{d}\| \le \frac{|\lambda_1|}{\lambda_2}\|d\|. \tag{51}$$

Proof We apply Lemma 3 with $p_1 = d$, $p_2 = \hat{d}$, and $y = k$ as given by (50). For $d \neq 0$ it clearly holds $\|\hat{d}\| \le \frac{|\lambda_1|}{\lambda_2}\|d\| < \|d\|$ as required. We have to verify

$$1 \le t = \frac{\gamma_2}{\gamma_2 - \lambda_1} \le 2t_{\min} - 1. \tag{52}$$

Because $\gamma_2 \ge \lambda_2 \ge \lambda_1$, the first inequality is clear. For the second we will show that even

$$t = \frac{\gamma_2}{\gamma_2-\lambda_1} \le t_{\min} = \frac{\langle d, d-\hat{d}\rangle}{\|d-\hat{d}\|^2} \quad \big(\le 2t_{\min} - 1\big) \tag{53}$$

holds, which by elementary manipulations is equivalent to

$$\gamma_2 \ge \lambda_1 \frac{\langle d, d-\hat{d}\rangle}{\langle \hat{d}, d-\hat{d}\rangle},$$

where we use $\langle d, d-\hat{d}\rangle = \sum_{i=2}^n (1-\frac{\lambda_1}{\lambda_i})d_i^2 > 0$ and $\frac{1}{\lambda_1}\langle \hat{d}, d-\hat{d}\rangle = \sum_{i=2}^n \frac{1}{\lambda_i}(1-\frac{1}{\lambda_i})d_i^2 > 0$, where the d_i are the components of $d = (d_2, \ldots, d_n)^T$. With γ_2 given by (49) we obtain the following sequence of inequalities, all equivalent to (53):

$$\begin{aligned} \langle d, d-\hat{d}\rangle \|d-\hat{d}\|^2 &\le \frac{1}{\lambda_1}\langle d-\hat{d}, \Lambda_2(d-\hat{d})\rangle \langle \hat{d}, d-\hat{d}\rangle, \\ \langle d, d-\hat{d}\rangle^2 - \langle d, d-\hat{d}\rangle\langle \hat{d}, d-\hat{d}\rangle &\le \frac{1}{\lambda_1}\langle \Lambda_2 d, d-\hat{d}\rangle\langle \hat{d}, d-\hat{d}\rangle - \langle d, d-\hat{d}\rangle\langle \hat{d}, d-\hat{d}\rangle, \\ \langle d, d-\hat{d}\rangle^2 &\le \left\langle \frac{1}{\lambda_1}\Lambda_2 d, d-\hat{d}\right\rangle \langle \hat{d}, d-\hat{d}\rangle. \end{aligned}$$

Here the last inequality can be written as

$$\left(\sum_{i=2}^{n}\left(1-\frac{\lambda_1}{\lambda_i}\right)d_i^2\right)^2 \;\le\; \left(\sum_{i=2}^{n}\lambda_i\left(1-\frac{\lambda_1}{\lambda_i}\right)d_i^2\right)\left(\sum_{i=2}^{n}\frac{1}{\lambda_i}\left(1-\frac{\lambda_1}{\lambda_i}\right)d_i^2\right)$$

which holds by the Cauchy–Schwarz inequality. This proves (53) and establishes Proposition 3 for $m = 1$. □

A.4 The Special Case $n = 2m + 1$

For $n = 2m + 1$ there exists a vector $w_0 \in \mathbb{R}^m$ satisfying $\|w_0\| = 1$ and $W^T w_0 = 0$, where $W \in \mathbb{R}^{(m+1)\times m}$ is given in Proposition 2. These conditions determine w_0 uniquely up to sign. Augmenting matrix W by the column vector w_0 yields an orthogonal matrix, which (for convenience) we denote by V^T,

$$V^T := [W w_0], \quad V^T V = V V^T = I_{m+1}.$$

From $W^T \Lambda_2 W = \Gamma_2$ it follows

$$V \Lambda_2 V^T = \left[\begin{array}{c|c} \Gamma_2 & y \\ \hline y^T & \sigma \end{array}\right] \tag{54}$$

with

$$y = W^T \Lambda_2 w_0 \quad \text{and} \quad \sigma = w_0^T \Lambda_2 w_0.$$

γ_j satisfy

$$\lambda_{m+1} \le \gamma_{m+1} \le \lambda_{m+2} \le \gamma_{m+2} \le \cdots \le \lambda_{2m} \le \gamma_{2m} \le \lambda_{2m+1}, \tag{55}$$

cf. [9, Theorem 4.3.8].

[9, Theorem 4.3.10] gives the converse of the statement (55): For a given sequence $\gamma_{m+1}, \ldots, \gamma_{2m}$ of real numbers which satisfy (55) there exist essentially unique $\sigma \in \mathbb{R}$ and $y \in \mathbb{R}^m$ (σ is unique while the components of y are unique up to sign) such that the symmetric matrix on the right of (54) has eigenvalues $\lambda_{m+1}, \ldots, \lambda_{2m+1}$, i.e., such that (54) holds after a diagonalization of the matrix. σ and $y = (y_{m+1}, \ldots, y_{2m})^T$ are explicitly given by

$$\begin{aligned} \sigma &= \text{trace } V \Lambda_2 V^T - \text{trace } \Gamma_2 \;=\; \sum_{i=m+1}^{2m+1}\lambda_i - \sum_{j=m+1}^{2m}\gamma_j, \\ y_j^2 &= -\frac{\prod_{i=m+1}^{2m+1}(\gamma_j - \lambda_i)}{\prod_{\substack{i=m+1 \\ i\ne j}}^{2m}(\gamma_j - \gamma_i)}, \quad j = m+1, \ldots, 2m. \end{aligned} \tag{56}$$

Let $v_i = \left[\begin{array}{c} z_i \\ \hline w_{0i} \end{array}\right]$ with $z_i \in \mathbb{R}^m$, $i = m+1, \ldots, 2m+1$ be the columns of the matrix V. v_i is an eigenvector of norm 1 of the matrix (54) corresponding to the eigenvalue λ_i, thus

$$\left(\left[\begin{array}{c|c} \Gamma_2 & y \\ \hline y^T & \sigma \end{array}\right] - \lambda_i I_{m+1}\right) v = \left[\begin{array}{c|c} \Gamma_2 - \lambda_i I_m & y \\ \hline y^T & \sigma - \lambda_i \end{array}\right] \left[\begin{array}{c} z_i \\ \hline w_{0i} \end{array}\right] = 0,$$

whence

$$z_i = -w_{0i}(\Gamma_2 - \lambda_i I_m)^{-1} y.$$

The requirement $\|v_i\|^2 = w_{0i}^2 + \|z_i\|^2 = 1$ gives

$$\left(1 + \sum_{j=m+1}^{2m}\frac{y_j^2}{(\gamma_j - \lambda_i)^2}\right) w_{0i}^2 = 1,$$

which after inserting (56) leads to

$$w_{0i}^2 \;=\; \frac{\prod_{j=m+1}^{2m}(\gamma_j-\lambda_i)}{\prod_{\substack{j=m+1\\ j\neq i}}^{2m+1}(\lambda_i-\lambda_j)}, \qquad i=m+1,\dots,2m+1. \tag{57}$$

Let d_j, $j=1,\dots,m$ denote the columns of D_2, and let $\hat{d}_j = \lambda_j \Lambda_2^{-1} d_j$, $j=1,\dots,m$ denote the columns of $\hat{D}_2$, cf. Proposition 1. By definition of W, $d_j - \hat{d}_j$ lies in the span of W, so there exists an $x_j \in \mathbb{R}^m$ with $(I_{m+1} - \lambda_j \Lambda_2^{-1}) d_j = d_j - \hat{d}_j = W x_j$, or

$$d_j = \Lambda_2(\Lambda_2 - \lambda_j I_{m+1})^{-1} W x_j, \quad j=1,\dots,m.$$

Similarly,

$$\begin{aligned} k_j &= \hat{d}_j - \lambda_j W(\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j) \\ &= \lambda_j \Lambda_2^{-1} \Lambda_2 (\Lambda_2 - \lambda_j I_{m+1})^{-1} W x_j - \lambda_j W(\Gamma_2 - \lambda_j I_m)^{-1} W^T W x_j \\ &= \lambda_j \big((\Lambda_2 - \lambda_j I_{m+1})^{-1} - W(\Gamma_2 - \lambda_j I_m)^{-1} W^T \big) W x_j, \quad j=1,\dots,m. \end{aligned}$$

With

$$A_j \;=\; (\Lambda_2 - \lambda_j I_{m+1})^{-1} - W(\Gamma_2 - \lambda_j I_m)^{-1} W^T, \tag{58}$$

$$B_j \;=\; \Lambda_2(\Lambda_2 - \lambda_j I_{m+1})^{-1}, \quad j=1,\dots,m \tag{59}$$

we have $d_j = B_j W x_j$ and $k_j = \lambda_j A_j W x_j$, therefore

$$\|k_j\| \;\le\; |\lambda_j| \max_{x\neq 0} \frac{\|A_j W x\|}{\|B_j W x\|} \|d_j\| \tag{60}$$

$$\;=\; |\lambda_j| \max_{y \in \{Wx : x\in\mathbb{R}^m\}\setminus\{0\}} \frac{\|A_j y\|}{\|B_j y\|} \|d_j\| \tag{61}$$

$$\;\le\; |\lambda_j| \max_{y\neq 0} \frac{\|A_j y\|}{\|B_j y\|} \|d_j\|, \quad j=1,\dots,m. \tag{62}$$

We rely for our further analysis on the latter estimate (62) which is based on the matrices A_j and B_j, since these are easier to handle than the matrices $A_j W$ and $B_j W$ on which the sharper estimate (60) is based. It turns out that A_j has a very simple structure: As

$$\begin{aligned} W^T(\Lambda_2 - \lambda_j I_{m+1}) A_j &= W^T - W^T(\Lambda_2 - \lambda_j I_{m+1}) W (\Gamma_2 - \lambda_j I_m)^{-1} W^T \\ &= W^T - (\Gamma_2 - \lambda_j I_m)(\Gamma_2 - \lambda_j I_m)^{-1} W^T \\ &= 0, \end{aligned}$$

the range of $(\Lambda_2 - \lambda_j I_{m+1}) A_j (\Lambda_2 - \lambda_j I_{m+1})$ is orthogonal to W and thus must be equal to the span of w_0. It follows that $(\Lambda_2 - \lambda_j I_{m+1}) A_j (\Lambda_2 - \lambda_j I_{m+1})$ is a symmetric rank-one matrix of the form $\alpha_j w_0 w_0^T$, and thus A_j itself is a symmetric rank-one matrix of the form

$$A_j = \alpha_j a_j a_j^T \quad \text{with} \quad a_j = (\Lambda_2 - \lambda_j I_{m+1})^{-1} w_0.$$

Using $W^T w_0 = 0$ we obtain from (58)

$$w_0^T A_j w_0 = w_0^T (\Lambda_2 - \lambda_j I_{m+1})^{-1} w_0.$$

On the other hand,

$$w_0^T A_j w_0 = \alpha_j w_0^T a_j a_j^T w_0 = \alpha_j w_0^T (\Lambda_2 - \lambda_j I_{m+1})^{-1} w_0 w_0^T (\Lambda_2 - \lambda_j I_{m+1})^{-1} w_0$$

whence

$$\alpha_j = \frac{1}{w_0^T (\Lambda_2 - \lambda_j I_{m+1})^{-1} w_0} = 1 \Big/ \sum_{i=m+1}^{2m+1} \frac{w_{0i}^2}{\lambda_i - \lambda_j}\,.$$

Inserting (57) this yields

$$\alpha_j = \frac{\prod_{i=m+1}^{2m+1}(\lambda_i - \lambda_j)}{\prod_{i=m+1}^{2m}(\gamma_i - \lambda_j)}, \quad j = 1, \ldots, m.$$

The maximum value of the bound (62) is equal to the square root of the largest eigenvalue λ of the generalized eigenvalue problem

$$A_j^2 x = \lambda B_j^2 x,$$

or, equivalently, of the standard eigenvalue problem

$$B_j^{-1} A_j^2 B_j^{-1} x = \lambda x, \tag{63}$$

whose matrix is again a symmetric rank-one matrix

$$B_j^{-1} A_j^2 B_j^{-1} = \alpha_j^2 \|a_j\|^2 b_j b_j^T \quad \text{with} \quad b_j = B_j^{-1} a_j = \Lambda_2^{-1} w_0.$$

Now it is easily seen that b_j is an eigenvector of (63) corresponding to the eigenvalue $\alpha_j^2 \|a_j\|^2 \|b_j\|^2$, which finally implies

$$\begin{aligned}
\|k_j\| &\leq |\lambda_j| \max_{y \neq 0} \frac{\|A_j y\|}{\|B_j y\|} \|d_j\| \\
&= |\lambda_j| |\alpha_j| \|a_j\| \|b_j\| \|d_j\| \\
&= |\lambda_j| |\alpha_j| \|(\Lambda_2 - \lambda_j I_{m+1})^{-1} w_0\| \|\Lambda_2^{-1} w_0\| \|d_j\| \\
&\leq |\lambda_j| \sqrt{\max_{(\gamma_{m+1}, \ldots, \gamma_{2m}) \in K} f(\gamma_{m+1}, \ldots, \gamma_{2m}; \lambda_j; \lambda_{m+1}, \ldots, \lambda_{2m+1})} \|d_j\|.
\end{aligned} \tag{64}$$

Here $K = K(\lambda_{m+1}, \ldots, \lambda_{2m+1})$ denotes the rectangle

$$\begin{aligned}
K = \{&(\gamma_{m+1}, \ldots, \gamma_{2m}) \in \mathbb{R}^m : \\
&\lambda_{m+1} \leq \gamma_{m+1} \leq \lambda_{m+2} \leq \gamma_{m+2} \leq \cdots \leq \lambda_{2m} \leq \gamma_{2m} \leq \lambda_{2m+1}\}
\end{aligned}$$

and the function $f(\gamma_{m+1}, \ldots, \gamma_{2m}) = f(\gamma_{m+1}, \ldots, \gamma_{2m}; \lambda_j; \lambda_{m+1}, \ldots, \lambda_{2m+1})$ is defined as

$$\begin{aligned}
f(\gamma_{m+1}, \ldots, \gamma_{2m}) &= f(\gamma_{m+1}, \ldots, \gamma_{2m}; \lambda_j; \lambda_{m+1}, \ldots, \lambda_{2m+1}) \\
&= \alpha_j^2 \|(\Lambda_2 - \lambda_j I_{m+1})^{-1} w_0\|^2 \|\Lambda_2^{-1} w_0\|^2 \\
&= \frac{\prod_{l=m+1}^{2m+1}(\lambda_l - \lambda_j)^2}{\prod_{l=m+1}^{2m}(\gamma_l - \lambda_j)^2} \left(\sum_{i=m+1}^{2m+1} \frac{1}{(\lambda_i - \lambda_j)^2} \frac{\prod_{l=m+1}^{2m}(\gamma_l - \lambda_i)}{\prod_{\substack{l=m+1 \\ l \neq i}}^{2m+1}(\lambda_l - \lambda_i)} \right) \\
&\qquad \times \left(\sum_{i=m+1}^{2m+1} \frac{1}{\lambda_i^2} \frac{\prod_{l=m+1}^{2m}(\gamma_l - \lambda_i)}{\prod_{\substack{l=m+1 \\ l \neq i}}^{2m+1}(\lambda_l - \lambda_i)} \right).
\end{aligned} \tag{65}$$

Finally, by applying Lemma 4 below, Proposition 3 for the special case $n = 2m + 1$ follows from (64).

Lemma 4 *Let $0 < \lambda_1 < \cdots < \lambda_{m+1}$ and $\omega < \lambda_1$ be given. Let $K = K(\lambda_1, \ldots, \lambda_{m+1}) \subseteq \mathbb{R}^m$ denote the rectangle*

$$K = \{(\gamma_1, \ldots, \gamma_m) \in \mathbb{R}^m : \ \lambda_1 \leq \gamma_1 \leq \lambda_2 \leq \gamma_2 \leq \cdots \leq \lambda_m \leq \gamma_m \leq \lambda_{m+1}\}, \tag{66}$$

and let $f : \ K \to \mathbb{R}$ be the function given by

$$f(\gamma_1, \ldots, \gamma_m) = \frac{f_A(\gamma_1, \ldots, \gamma_m) f_B(\gamma_1, \ldots, \gamma_m)}{f_C(\gamma_1, \ldots, \gamma_m)^2}, \tag{67}$$

where

$$f_A(\gamma_1,\dots,\gamma_m) = \sum_{i=1}^{m+1} \frac{1}{(\lambda_i-\omega)^2} \frac{\prod_{j=1}^{m}(\gamma_j-\lambda_i)}{\prod_{\substack{j=1\\ j\neq i}}^{m+1}(\lambda_j-\lambda_i)}, \tag{68}$$

$$f_B(\gamma_1,\dots,\gamma_m) = \sum_{i=1}^{m+1} \frac{1}{\lambda_i^2} \frac{\prod_{j=1}^{m}(\gamma_j-\lambda_i)}{\prod_{\substack{j=1\\ j\neq i}}^{m+1}(\lambda_j-\lambda_i)}, \tag{69}$$

$$f_C(\gamma_1,\dots,\gamma_m) = \frac{\prod_{i=1}^{m}(\gamma_i-\omega)}{\prod_{i=1}^{m+1}(\lambda_i-\omega)}. \tag{70}$$

Then it holds

$$\max_{(\gamma_1,\dots,\gamma_m)\in K} f(\gamma_1,\dots,\gamma_m) \leq \begin{cases} \dfrac{1}{\lambda_1^2} & \text{for} \quad \omega/\lambda_1 \leq 2\sqrt{2}-2, \\ \dfrac{1}{16\omega^2}\dfrac{(2-\omega/\lambda_1)^4}{1-\omega/\lambda_1} & \text{for} \quad \omega/\lambda_1 \geq 2\sqrt{2}-2. \end{cases} \tag{71}$$

Proof First, note that from $\omega < \lambda_1 \leq \gamma_1 \leq \lambda_2 \leq \gamma_2 \leq \cdots \leq \lambda_m \leq \gamma_m \leq \lambda_{m+1}$ it follows easily

$$f(\gamma_1,\dots,\gamma_m) \geq 0 \quad \text{for } (\gamma_1,\dots,\gamma_m) \in K. \tag{72}$$

Next, we show that if $m \geq 2$, then f has no local maximum in the interior of K. The factor $\prod_{j=1}^{m}(\omega-\gamma_j) = (-1)^m \prod_{j=1}^{m}(\gamma_j-\omega)$ occurring in f_C can be written as

$$\prod_{j=1}^{m}(\omega-\gamma_j) = \omega^m - s_1\omega^{m-1} + s_2\omega^{m-2} - \cdots + (-1)^m s_m,$$

where s_j are the elementary symmetric polynomials in the variables $\gamma_1,\dots,\gamma_m$,

$$\begin{aligned} s_1 &= \gamma_1 + \cdots \gamma_m, \\ s_2 &= \gamma_1\gamma_2 + \gamma_1\gamma_3 + \cdots + \gamma_{m-1}\gamma_m = \sum_{1\leq i<j\leq m} \gamma_i\gamma_j, \\ s_3 &= \sum_{1\leq i<j<k\leq m} \gamma_i\gamma_j\gamma_k, \\ &\vdots \\ s_m &= \gamma_1\gamma_2\cdots\gamma_m. \end{aligned}$$

It follows that f_C is a *linear* function of $s_1,\dots,s_m$,

$$f_C(\gamma_1,\dots,\gamma_m) = \tilde{f}_C(s_1,\dots,s_m) = c_0 + \sum_{j=1}^{m} c_j s_j \tag{73}$$

with coefficients c_j depending on ω and $\lambda_1,\dots,\lambda_{m+1}$. Similarly by expanding

$$\prod_{j=1}^{m}(\lambda_i-\gamma_j) = \lambda_i^m - s_1\lambda_i^{m-1} + s_2\lambda_i^{m-1} - \cdots + (-1)^m s_m, \quad i = 1,\dots,m+1$$

it follows that also f_A and f_B are *linear* functions of $s_1,\dots,s_m$,

$$\tilde{f}_A(s_1,\dots,s_m) = a_0 + \sum_{j=1}^{m} a_j s_j, \quad \tilde{f}_B(s_1,\dots,s_m) = b_0 + \sum_{j=1}^{m} b_j s_j, \tag{74}$$

where the coefficients a_j, b_j again depend on ω and $\lambda_1, \ldots, \lambda_{m+1}$. In summary, the variable transformation

$$F: \ (\gamma_1, \ldots, \gamma_m) \mapsto (s_1, \ldots, s_m)$$

gives $f(\gamma_1, \ldots, \gamma_m)$ in the form

$$f(\gamma_1, \ldots, \gamma_m) = \tilde{f}(s_1, \ldots, s_m) = \frac{\tilde{f}_A(s_1, \ldots, s_m)\tilde{f}_B(s_1, \ldots, s_m)}{\tilde{f}_C(s_1, \ldots, s_m)^2},$$

where each of $\tilde{f}_A, \tilde{f}_B, \tilde{f}_C$ is a *linear* function of $s_1, \ldots, s_m$, cf. (74), (73). The Jacobian determinant of F is given by

$$\det DF(\gamma_1, \ldots, \gamma_m) = \prod_{1 \le i < j \le m} (\gamma_i - \gamma_j)$$

(cf. [14]), which vanishes if and only if $\gamma_j = \gamma_{j+1} = \lambda_{j+1}$ for some $1 \le j \le m-1$, i.e., on the boundary of K. It follows that $\det DF(\gamma_1, \ldots, \gamma_m) \neq 0$ for $(\gamma_1, \ldots, \gamma_m)$ in the *interior* of K. Thus a local maximum $(\gamma_{*1}, \ldots, \gamma_{*m})$ of f in the interior of K would be a critical point $(s_{*1}, \ldots, s_{*m}) = F(\gamma_{*1}, \ldots, \gamma_{*m})$ of $\tilde{f}$. Lemma 5 below would thus imply $f(\gamma_{*1}, \ldots, \gamma_{*m}) = \tilde{f}(s_{*1}, \ldots, s_{*m}) = 0$, which cannot hold for a local maximum of f in the interior of K, since f is nonnegative there, cf. (72). This shows that f has no local maximum in the interior of K and thus the maximum of f is attained at the boundary of K.

To find the maximum of f on K, we have to examine f on the boundary of K, so let us assume that $\gamma_k = \lambda_{k_*}$ holds for some $1 \le k \le m$ and $k_* = k$ or $k_* = k+1$. Then, due to $\prod_{j=1}^m (\gamma_j - \lambda_{k_*}) = 0$, it follows from (68)–(70)

$$f_A(\gamma_1, \ldots, \gamma_m) = \sum_{\substack{i=1 \\ i \neq \lambda_{k_*}}}^{m+1} \frac{1}{(\lambda_i - \omega)^2} \underbrace{\frac{\gamma_k - \lambda_i}{\lambda_{k_*} - \lambda_i}}_{=1} \frac{\prod_{\substack{j=1 \\ j \neq k}}^m (\gamma_j - \lambda_i)}{\prod_{\substack{j=1 \\ j \neq i, j \neq k_*}}^{m+1} (\lambda_j - \lambda_i)}, \tag{75}$$

$$f_B(\gamma_1, \ldots, \gamma_m) = \sum_{\substack{i=1 \\ i \neq \lambda_{k_*}}}^{m+1} \frac{1}{\lambda_i^2} \underbrace{\frac{\gamma_k - \lambda_i}{\lambda_{k_*} - \lambda_i}}_{=1} \frac{\prod_{\substack{j=1 \\ j \neq k}}^m (\gamma_j - \lambda_i)}{\prod_{\substack{j=1 \\ j \neq i, j \neq k_*}}^{m+1} (\lambda_j - \lambda_i)}, \tag{76}$$

$$f_C(\gamma_1, \ldots, \gamma_m) = \underbrace{\frac{\gamma_k - \omega}{\lambda_{k_*} - \omega}}_{=1} \frac{\prod_{\substack{i=1 \\ i \neq k_*}}^m (\gamma_i - \omega)}{\prod_{i=1}^{m+1} (\lambda_i - \omega)}. \tag{77}$$

Thus, if the value $\gamma_k = \lambda_{k_*}$ is fixed, then γ_k and λ_{k_*} no longer appear in (75)–(77), and after renumbering $\gamma_{j+1} \to \gamma_j$, $j = k, \ldots, m-1$ and similarly $\lambda_{j+1} \to \lambda_j$, $j = k_*, \ldots, m$ we obtain the formulae corresponding to (67)–(70) for the $m-1$ variables $\gamma_1, \ldots, \gamma_{m-1}$ instead of m. Note that after removing γ_k, λ_{k_*} and renumbering the remaining variables as stated, the constraints

$$\omega < \lambda_1 \le \gamma_1 \le \lambda_2 \le \cdots \le \lambda_{m-1} \le \gamma_{m-1} \le \lambda_m$$

remain valid. Furthermore, the bound (71) is monotonically decreasing with respect to λ_1. Therefore a bound of the form (71) with the original (instead of the possibly renumbered) λ_1 for $f(\gamma_1, \ldots, \gamma_{m-1})$ is also a bound for $f(\gamma_1, \ldots, \gamma_m)$. This reduction $m \to m-1$ of the problem by the above argument can be repeated as long as $m \ge 2$, and it follows that if (71) holds in the special case $m = 1$, then it holds for all $m \ge 1$.

It remains to prove (71) in the special case $m = 1$. Thus,

$$f(\gamma_1) = \frac{(\lambda_1\lambda_2 - (\lambda_1 + \lambda_2)\gamma_1)(\lambda_1\lambda_2 - \omega^2 + (2\omega - \lambda_1 - \lambda_2)\gamma_1)}{\lambda_1^2\lambda_2^2(-\omega + \gamma_1)^2}.$$

The maximum of $f(\gamma_1)$ for $\gamma_1 \in [\lambda_1, \lambda_2]$ is either attained at the boundary where it holds

$$f(\lambda_1) = \frac{1}{\lambda_2^2} \le \frac{1}{\lambda_1^2}, \qquad f(\lambda_2) = \frac{1}{\lambda_1^2}, \tag{78}$$

or the maximum is attained at a critical point $\gamma_{*1} \in (\lambda_1, \lambda_2)$. Solving $\frac{d}{d\gamma_1} f(\gamma_1) = 0$ yields a unique critical point

$$\gamma_{*1} = \frac{2(\lambda_1+\lambda_2)^2 + \omega^3(\lambda_1+\lambda_2) - 2\omega\lambda_1\lambda_2(\lambda_1+\lambda_2)}{2\lambda_1\lambda_2(\lambda_1+\lambda_2) + 3\omega^2(\lambda_1+\lambda_2) - 2\omega(\lambda_1+\lambda_2)^2 - 2\omega\lambda_1\lambda_2},$$

where f has the value

$$f(\gamma_{*1}) = \frac{\omega^2(2\lambda_1\lambda_2 - \omega(\lambda_1+\lambda_2))^2}{4\lambda_1^2\lambda_2^2(\lambda_1-\omega)(\lambda_2-\omega)(\omega(\lambda_1+\lambda_2)-\lambda_1\lambda_2)}.$$

Let us consider the case that $\gamma_{*1} \in (\lambda_1, \lambda_2)$ and that the maximal value of $f(\gamma_1)$ for $\gamma_1 \in [\lambda_1, \lambda_2]$ is given by $f(\gamma_{*1})$. (Otherwise the maximum would be attained at $\gamma_1 = \lambda_1$ or $\gamma_1 = \lambda_2$ and (78) would apply.) We regard the expression for $f(\gamma_{*1})$ as a (rational) function of λ_2,

$$g(\lambda_2) = \frac{\omega^2(2\lambda_1\lambda_2 - \omega(\lambda_1+\lambda_2))^2}{4\lambda_1^2\lambda_2^2(\lambda_1-\omega)(\lambda_2-\omega)(\omega(\lambda_1+\lambda_2)-\lambda_1\lambda_2)},$$

and want to find the maximum of $g(\lambda_2)$ for $\lambda_2 \in [\gamma_{*1}, \infty)$. The existence of this maximum follows from $\lim_{\lambda_2\to\infty} g(\lambda_2) = 0$ and the fact that g has no poles in $[\gamma_{*1}, \infty)$. Note that $\omega(\lambda_1+\lambda_2) \neq \lambda_1\lambda_2$ because otherwise $\gamma_{*1} = \omega$ contradicting our assumption $\omega < \lambda_1 < \gamma_{*1}$. The maximum has to be attained either at the boundary $\lambda_2 = \gamma_{*1}$, where it holds $g(\gamma_{*1}) = f(\lambda_2) = 1/\lambda_1^2$, or at one of the critical points of g, i.e., at one of the zeros of

$$\frac{d}{d\lambda_2} g(\lambda_2) = \frac{\omega^2(2\lambda_1\lambda_2 - \omega(\lambda_1+\lambda_2))(2\lambda_1\lambda_2 - \omega(2\lambda_1+\lambda_2))(2\lambda_1\lambda_2^2 + \omega^2(\lambda_1+\lambda_2) - 2\omega\lambda_2(\lambda_1+\lambda_2))}{4\lambda_1^2\lambda_2^3(\lambda_1-\omega)(\lambda_2-\omega)^2(\lambda_1\lambda_2 - \omega(\lambda_1+\lambda_2))^2}.$$

If the critical point λ_2 is a zero of the factor $(2\lambda_1\lambda_2^2 + \omega^2(\lambda_1+\lambda_2) - 2\omega\lambda_2(\lambda_1+\lambda_2))$ of the numerator of $\frac{d}{d\lambda_2} g(\lambda_2)$, then it is also a zero of

$$\gamma_{*1} - \lambda_2 = \frac{(\lambda_2-\omega)(2\lambda_1\lambda_2^2 + \omega^2(\lambda_1+\lambda_2) - 2\omega\lambda_2(\lambda_1+\lambda_2))}{2\lambda_1\lambda_2(\lambda_1+\lambda_2) + 3\omega^2(\lambda_1+\lambda) - 2\omega(\lambda_1+\lambda_2)^2 - 2\omega\lambda_1\lambda_2},$$

so that $\lambda_2 = \gamma_{*1}$ and thus $g(\lambda_2) = g(\gamma_{*1}) = f(\lambda_2) = 1/\lambda_1^2$. If on the other hand λ_2 is a zero of the factor $(2\lambda_1\lambda_2 - \omega(\lambda_1+\lambda_2))$, then $g(\lambda_2) = 0$ which is not the maximum of g on $[\gamma_{*1}, \infty)$. It remains to discuss the case that the critical point λ_2 is a zero of the factor $(2\lambda_1\lambda_2 - \omega(2\lambda_1+\lambda_2))$. In this case it follows

$$\lambda_2 = \frac{2\omega\lambda_1}{2\lambda_1-\omega} \quad \text{with} \quad g(\lambda_2) = \frac{1}{16}\frac{(2\lambda_1-\omega)^4}{\omega^2\lambda_1^3(\lambda_1-\omega)} = \frac{1}{16\omega^2}\frac{(2-\omega/\lambda_1)^4}{1-\omega/\lambda_1},$$

which is a possible candidate for the maximum of g on $[\gamma_{*1}, \infty)$. Summarizing, we have thus proven

$$\begin{aligned}
\max_{\gamma\in[\lambda_1,\lambda_2]} f(\gamma) &\le \max\Big\{\frac{1}{\lambda_1}, f(\gamma_{*1})\Big\} \\
&\le \max\Big\{\frac{1}{\lambda_1}, \max_{\lambda_2\in[\gamma_{*1},\infty)} g(\lambda_2)\Big\} \\
&= \max\Big\{\frac{1}{\lambda_1}, \frac{1}{16\omega^2}\frac{(2-\omega/\lambda_1)^4}{1-\omega/\lambda_1}\Big\} \\
&= \begin{cases} \dfrac{1}{\lambda_1^2} & \text{for } \omega/\lambda_1 \le 2\sqrt{2}-2, \\ \dfrac{1}{16\omega^2}\dfrac{(2-\omega/\lambda_1)^4}{1-\omega/\lambda_1} & \text{for } \omega/\lambda_1 \ge 2\sqrt{2}-2, \end{cases}
\end{aligned}$$

which is the proposition of the lemma in the special case $m = 1$. As shown before, the cases $2 \le m \le 2n+1$ follow from this result. □

Lemma 5 *Consider a function* $f:\ \Omega \subseteq \mathbb{R}^m \to \mathbb{R}$, $m \geq 2$ *of the form*

$$f(x_1,\ldots,x_m) = \frac{(a_0 + a_1x_1 + \cdots + a_mx_m)(b_0 + b_1x_1 + \cdots + b_mx_m)}{(c_0 + c_1x_1 + \cdots + c_mx_m)^2},$$

where $a_j, b_j, c_j \in \mathbb{R}$, $j = 0,\ldots,m$ *and* $a_1^2+\cdots+a_m^2 \neq 0$, $b_1^2+\cdots+b_m^2 \neq 0$, $c_1^2+\cdots+c_m^2 \neq 0$, *and where* $\Omega \subseteq \mathbb{R}^m$ *denotes an open domain in which* $c_0 + c_1x_1 + \cdots + c_mx_m \neq 0$. *Let* $(x_{*1},\ldots,x_{*m}) \in \Omega$ *be a critical point of* f, *i.e.,*

$$\frac{\partial}{\partial x_i} f(x_1,\ldots,x_m)\Big|_{(x_1,\ldots,x_m)=(x_{*1},\ldots,x_{*m})} = 0, \quad i = 1,\ldots,m.$$

Then it holds

$$f(x_{*1},\ldots,x_{*m}) = 0. \tag{79}$$

Proof We first prove the lemma in the special case $m = 2$ and then show that the general case is a direct consequence. Let $m = 2$. We consider the two cases (i) $\det \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} \neq 0$ and (ii) $\det \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} = 0$.

In the case (i) we apply the variable transformation $(x_1, x_2) \mapsto (u_1, u_2) = (a_0 + a_1x_1 + a_2x_2, b_0 + b_1x_1 + b_2x_2)$ whose Jacobian determinant is a constant $\neq 0$. Thereby f is transplanted to a function $\tilde{f}$ of the form

$$\tilde{f}(u_1,u_2) = \frac{u_1u_2}{(\tilde{c}_0 + \tilde{c}_1u_1 + \tilde{c}_2u_2)^2}, \quad \tilde{c}_1^2 + \tilde{c}_2^2 \neq 0,$$

whose derivatives are given by

$$\frac{\partial}{\partial u_1}\tilde{f}(u_1,u_2) = \frac{u_2(\tilde{c}_0 - \tilde{c}_1u_1 + \tilde{c}_2u_2)}{(\tilde{c}_0 + \tilde{c}_1u_1 + \tilde{c}_2u_2)^3}, \quad \frac{\partial}{\partial u_2}\tilde{f}(u_1,u_2) = \frac{u_1(\tilde{c}_0 + \tilde{c}_1u_1 - \tilde{c}_2u_2)}{(\tilde{c}_0 + \tilde{c}_1u_1 + \tilde{c}_2u_2)^3}.$$

A straightforward calculation shows that the only solution (u_1, u_2) of

$$\begin{cases} u_2(\tilde{c}_0 - \tilde{c}_1u_1 + \tilde{c}_2u_2) = 0, \\ u_1(\tilde{c}_0 + \tilde{c}_1u_1 - \tilde{c}_2u_2) = 0, \\ \quad \tilde{c}_0 + \tilde{c}_1u_1 + \tilde{c}_2u_2 \neq 0 \end{cases} \tag{80}$$

is given by $(u_1, u_2) = (0, 0)$ provided that $\tilde{c}_0 \neq 0$. If $\tilde{c}_0 = 0$ then (80) has no solution. Since a critical point $(x_{*1}, x_{*2}) \in \Omega$ of f is transformed into a critical point (u_{*1}, u_{*2}) of $\tilde{f}$, i.e., to a solution of (80), it follows that (x_{*1}, x_{*2}) maps to $(u_{*1}, u_{*2}) = (0, 0)$ and thus

$$f(x_{*1}, x_{*2}) = \tilde{f}(u_{*1}, u_{*2}) = \tilde{f}(0, 0) = 0.$$

In the case (ii) there exists a variable transformation $(x_1, x_2) \mapsto (u_1, u_2)$ with constant Jacobian determinant $\neq 0$ such that f is transplanted to a function $\tilde{f}$ of the form

$$\tilde{f}(u_1,u_2) = \frac{u_1(\tilde{b}_0 + \tilde{b}_1u_1)}{(\tilde{c}_0 + \tilde{c}_1u_1 + \tilde{c}_2u_2)^2}, \quad \tilde{b}_1 \neq 0,\ \tilde{c}_1^2 + \tilde{c}_2^2 \neq 0.$$

The derivatives of $\tilde{f}$ are now given by

$$\begin{aligned} \frac{\partial}{\partial u_1}\tilde{f}(u_1,u_2) &= \frac{\tilde{b}_0(\tilde{c}_0 - \tilde{c}_1u1 + \tilde{c}_2u_2) + 2\tilde{b}_1u_1(\tilde{c}_0 + \tilde{c}_2u_2)}{(\tilde{c}_0 + \tilde{c}_1u_1 + \tilde{c}_2u_2)^3}, \\ \frac{\partial}{\partial u_2}\tilde{f}(u_1,u_2) &= \frac{-2\tilde{c}_2u_1(\tilde{b}_0 + \tilde{b}_1u1)}{(\tilde{c}_0 + \tilde{c}_1u_1 + \tilde{c}_2u_2)^3}. \end{aligned}$$

A straightforward calculation shows that

$$\begin{cases} \tilde{b}_0(\tilde{c}_0 - \tilde{c}_1 u_1 + \tilde{c}_2 u_2) + 2\tilde{b}_1 u_1(\tilde{c}_0 + \tilde{c}_2 u_2) = 0, \\ \tilde{c}_2 u_1(\tilde{b}_0 + \tilde{b}_1 u_1) = 0, \\ \tilde{c}_0 + \tilde{c}_1 u_1 + \tilde{c}_2 u_2 \neq 0 \end{cases} \tag{81}$$

has no solution so that in the case (ii) f does not have a critical point in Ω.

For the general case $m \geq 2$ write

$$f(x_1, \ldots, x_m) = \frac{(A_0(x_3, \ldots, x_m) + a_1 x_1 + a_2 x_2)(B_0(x_3, \ldots, x_m) + b_1 x_1 + b_2 x_2)}{(C_0(x_3, \ldots, x_m) + c_1 x_1 + c_2 x_2)^2},$$

where $A_0(x_3, \ldots, x_m) = a_0 + a_3 x_3 + \cdots + a_m x_m$ and B_0, C_0 are defined analogously. If $(x_{*1}, \ldots, x_{*m}) \in \Omega$ is a critical point of f then (x_{*1}, x_{*2}) is a critical point of g defined by

$$g(x_1, x_2) = f(x_1, x_2, x_{*3}, \ldots, x_{*m}) = \frac{(A_{*0} + a_1 x_1 + a_2 x_2)(B_{*0} + b_1 x_1 + b_2 x_2)}{(C_{*0} + c_1 x_1 + c_2 x_2)^2}, \tag{82}$$

where $A_{*0} = A_0(x_{*3}, \ldots, x_{*m})$ and B_{*0}, C_{*0} are defined similarly. By applying the special case $m = 2$ of the lemma to (82) it follows $g(x_{*1}, x_{*2}) = 0$ and thus (79), i.e., that the lemma holds in the general case $m \geq 2$ as well. □

A.5 The General Case

Let arbitrary $m \geq 1$ and $n \geq 2m + 1$ be given. If $n = 2m + 1$ then Proposition 3 holds, this was proven in Subsection A.4. So let us assume $n \geq 2m + 2$. For given $D_2 \in \mathbb{R}^{(n-m)\times m}$ and fixed j we apply Proposition 2 and obtain $W \in \mathbb{R}^{(n-m)\times m}$ and $\Gamma_2 = \mathrm{diag}(\gamma_{m+1}, \ldots, \gamma_{2m})$, such that

$$k_j = \hat{d}_j - \lambda_j W(\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j) \tag{83}$$

holds, where d_j denotes the j-th column of D_2, $\hat{d}_j = \lambda_j \Lambda_2^{-1} \hat{d}_j$, and k_j is the unique solution of the equation (the j-th column of the Sylvester equation (22))

$$P\Lambda_2 k_j - \lambda_j k_j + \lambda_j (I_{n-m} - P) d_j = 0 \tag{84}$$

with $P = WW^T$.

Let $\tilde{n} = n + 1$, $\tilde{m} = m + 1$. We form the matrix $\tilde{D}_2 = [D_2 \ \tilde{d}_{\tilde{m}}] \in \mathbb{R}^{(\tilde{n}-\tilde{m})\times\tilde{m}}$ by augmenting D_2 with some suitable column vector $\tilde{d}_{\tilde{m}} \in \mathbb{R}^{\tilde{n}-\tilde{m}}$ yet to be chosen. We choose $\tilde{\lambda}_{\tilde{m}} \in (\lambda_m, \lambda_{m+1})$ arbitrarily and define $\tilde{\Lambda}_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_m, \tilde{\lambda}_{\tilde{m}})$ and $\hat{\tilde{D}}_2 = \Lambda_2^{-1} \tilde{D}_2 \tilde{\Lambda}_1$. If we apply Proposition 2 with n, m, D_2, $\hat{D}_2$ respectively replaced by $\tilde{n}$, $\tilde{m}$, $\tilde{D}_2$, $\hat{\tilde{D}}_2$ (but with Λ_2 left unchanged) we obtain $\tilde{W} \in \mathbb{R}^{(\tilde{n}-\tilde{m})\times\tilde{m}}$ and $\tilde{\Gamma}_2 = \mathrm{diag}(\tilde{\gamma}_{\tilde{m}+1}, \ldots, \tilde{\gamma}_{2\tilde{m}})$ such that

$$\tilde{k}_j = \hat{d}_j - \lambda_j \tilde{W}(\tilde{\Gamma}_2 - \lambda_j I_{\tilde{m}})^{-1} \tilde{W}^T (d_j - \hat{d}_j) \tag{85}$$

holds, where now $\tilde{k}_j$ is the unique solution of the equation

$$\tilde{P}\Lambda_2 \tilde{k}_j - \lambda_j \tilde{k}_j + \lambda_j (I_{\tilde{n}-\tilde{m}} - \tilde{P}) d_j = 0 \tag{86}$$

with $\tilde{P} = \tilde{W}\tilde{W}^T$. Note that here d_j, $\hat{d}_j$, and λ_j are still the same as in (83), (84).

Our goal is to choose the vector $\tilde{d}_{\tilde{m}} \in \mathbb{R}^{\tilde{n}-\tilde{m}}$ in such a way that it holds $\tilde{k}_j = k_j$. If we assume that this is always possible and that it has already been proved that Proposition 3, eq. (47) is valid for n, m replaced by $\tilde{n}$, $\tilde{m}$, respectively, then due to $k_j = \tilde{k}_j$ the estimate (47) is valid for the particular k_j from (83), i.e. in this particular case also for the original values of n and m. (Note that λ_{m+1} occurring in (47) denotes the smallest eigenvalue of Λ_2 which remains unchanged in the transition $n \to \tilde{n}$, $m \to \tilde{m}$.) The argument above applies to any $D_2 \in \mathbb{R}^{(n-m)\times m}$ and index j and associated k_j. Therefore, if (47) holds for n, m replaced by $\tilde{n}$, $\tilde{m}$, respectively, then (47) holds also for the original values of n, m in general.

It is clear that this argument can be iterated. After $\ell = n - 2m - 1 \geq 1$ iterations we arrive at $n \to \tilde{\tilde{n}} = n + \ell$, $m \to \tilde{\tilde{m}} = m + \ell$ with $\tilde{\tilde{n}} = n + (n - 2m - 1) = 2(m + (n - 2m - 1)) + 1 = 2\tilde{\tilde{n}} + 1$, i.e., at the special case for which Proposition 3 has already been established, cf. Subsection A.4. Going backwards ($\tilde{\tilde{n}} \to n$, $\tilde{\tilde{m}} \to m$) it follows that Proposition 3 is valid for our originally chosen values of $m \geq 1$ and $n \geq 2m + 2$.

It remains to prove that the vector $\tilde{d}_{\tilde{m}} \in \mathbb{R}^{\tilde{n}-\tilde{m}}$ can always be chosen in such a way that it holds $\tilde{k} = k$, that is, such that k_j is the solution of equation (86), which depends on $\tilde{d}_{\tilde{m}}$ via $\tilde{P}$. We compute (cf. the proof of Proposition 2)

$$\begin{aligned}\tilde{P}\Lambda_2 k_j - \lambda_j k_j &= (\tilde{P}\Lambda_2 - \lambda_j I_{n-m}) \left(\hat{d}_j - \lambda_j W(\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j)\right) \\ &= \lambda_j \tilde{P} d_j - \lambda_j \hat{d}_j - \lambda_j \tilde{W}\tilde{W}^T \Lambda_2 W(\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j) \\ &\quad + \lambda_j^2 W(\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j), \end{aligned} \tag{87}$$

from which it follows that k_j is the solution of (86) if

$$\tilde{W}\tilde{W}^T \Lambda_2 W(\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j) = \lambda_j W(\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j).$$

Recalling that this equation holds with $\tilde{W}$ replaced by W, k_j solves (86) also when

$$(\tilde{P} - P)\Lambda_2 W(\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j) = 0, \tag{88}$$

where $P = WW^T$, $\tilde{P} = \tilde{W}\tilde{W}^T$. Note that P is the projection of $\mathbb{R}^{n-m}$ onto the space spanned by $D_2 - \hat{D}_2$, whereas $\tilde{P}$ is the projection of $\mathbb{R}^{\tilde{n}-\tilde{m}} = \mathbb{R}^{n-m}$ onto the larger space spanned by $\tilde{D}_2 - \hat{\tilde{D}}_2 = [D_2 - \hat{D}_2 \quad \tilde{d}_{\tilde{m}} - \hat{\tilde{d}}_{\tilde{m}}]$. If follows that for (88) to hold, we have to choose $\tilde{d}_{\tilde{m}}$ in such a way that $\tilde{d}_{\tilde{m}} - \hat{\tilde{d}}_{\tilde{m}} = (I_{n-m} - \tilde{\lambda}_{\tilde{m}} \Lambda_2^{-1}) \tilde{d}_{\tilde{m}}$ is orthogonal both to $\Lambda_2 W(\Gamma_2 - \lambda_j I_m)^{-1} W^T (d_j - \hat{d}_j)$ and to the space spanned by $D_2 - \hat{D}_2$, which clearly can be realized because

$$(\text{dimension of the space spanned by } D_2 - \hat{D}_2) + 1 = m + 1 < n - m$$

if $n \geq 2m + 2$.

This completes the proof of Proposition 3 and thus also of Theorem 1.

References

1. C. Bendtsen, O. Nielsen, and L. Hansen. Solving large nonlinear generalized eigenvalue problems from density functional theory calculations in parallel. *Appl. Numer. Math.*, 37:189–199, 2001.
2. P. Blaha, H. Hofstätter, O. Koch, R. Laskowsky, and K. Schwarz. Iterative diagonalization in APW based methods in electronic structure calculations. *J. Comput. Phys.*, 229:453–460, 2010.
3. P. Blaha, K. Schwarz, and G. Madsen. Electronic structure calculations of solids using the WIEN2k package for material sciences. *Comput. Phys. Commun.*, 147:71–76, 2002.
4. P. Blaha, K. Schwarz, G. Madsen, D. Kvasnicka, and J. Luitz. An augmented plane wave plus local orbital program for calculating crystal properties, 2001. ISBN 3-9501031-1-2.
5. M. Crouzeix, B. Philippe, and M. Sadkane. The Davidson method. *SIAM J. Sci. Comput.*, 15:62–76, 1994.
6. E. Davidson. The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices. *J. Comput. Phys.*, 17:87–94, 1975.
7. C.J. Garcia-Cervera, J. Lu, Y. Xuan, and W. E. Linear-scaling subspace-iteration algorithm with optimally localized nonorthogonal wave functions for Kohn-Sham density functional theory. *Phys. Rev. B*, 79:115110, 2009.
8. P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136(3B):B864, 1964.

9. R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.
10. W. Kohn and L.J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140(4A):A1133, 1965.
11. G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.*, 6:15–50, 1996.
12. G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54:11169–11186, 1996.
13. G. Kresse and J. Furthmüller. VASP the GUIDE, 2007. Available at `http://cms.mpi.univie.ac.at/VASP`.
14. A. Lascoux and P. Pragacz. Jacobians of symmetric polynomials. *Ann.Comb.*, 6:169–172, 2002.
15. R. Laskowski and P. Blaha. Unraveling the structure of the h-bn/rh(111) nanomesh with ab initio calculations. *J. Phys.: Condens. Matter*, 20:064207, 2008.
16. G.K.H. Madsen, P. Blaha, K. Schwarz, E. Sjöstedt, and L. Nordström. Efficient linearization of the augmented plane-wave method. *Phys. Rev. B*, 64:195134, 2001.
17. P. Pulay. Convergence acceleration of iterative sequences. The case of SCF iteration. *Chem. Phys. Lett.*, 73:393, 1980.
18. M.J. Rayson and P.R. Briddon. Rapid iterative method for electronic-structure eigenproblems using localized basis functions. *Comput. Phys. Commun.*, 178:128–134, 2008.
19. K. Schwarz. DFT calculations of solids with LAPW and WIEN2k. *Solid State Commun.*, 176:319–328, 2003.
20. K. Schwarz and P. Blaha. Solid state calculations using WIEN2k. *Comput. Mater. Sci.*, 28:259–273, 2003.
21. D. Singh. Simultaneous solution of diagonalization and self-consistency problems for transition-metal systems. *Phys. Rev. B*, 40:5428–5431, 1989.
22. G. L. G. Sleijpen, A. G. L. Booten, D. R. Fokkema, and H. A. Van der Vorst. Jacobi-Davidson type methods for generalized eigenproblems and polynomial eigenproblems. *BIT*, 36(3):595–633, 1996.
23. G. L. G. Sleijpen and H. A. Van der Vorst. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 17(2):401–425, 1996.
24. M. P. Teter, M. Payne, and D. C. Allan. Solution of Schrödinger's equation for large systems. *Phys. Rev. B*, 40(18):12255, 1989.
25. H. J. J. van Dam, J. H. van Lenthe, G. L. G. Sleijpen, and H. A. van der Vorst. An improvement of Davidson's iteration method. *J. Comput. Chem.*, 17, 3:267–272, 1996.
26. J. VandeVondele and J. Hutter. An efficient orbital transformation mathod for electronic structure calculations. *J. Chem. Phys.*, 118(10):4365–4369, 2003.
27. D.M. Wood and A. Zunger. A new method for diagonalising large matrices. *J. Phys. A: Math. Gen.*, 18:1343–1359, 1985.
28. Ch. Yang, W. Gao, and J.C. Meza. On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems. *SIAM. J. Matrix Anal. Appl.*, 30:1773–1788, 2009.
29. Y. Zhou, Y. Saad, M. Tiago, and J. Chelikowsky. Parallel self-consistent-field calculations via Chebyshev-filtered subspace acceleration. *Phys. Rev. E*, 74:066704, 2006.