A GEOMETRIC THEORY FOR PRECONDITIONED INVERSE ITERATION

I: EXTREMA OF THE RAYLEIGH QUOTIENT

KLAUS NEYMEYR

ABSTRACT. The discretization of eigenvalue problems for partial differential operators is a major source of matrix eigenvalue problems having very large dimensions, but only some of the smallest eigenvalues together with the eigenvectors are to be determined. Preconditioned inverse iteration (a "matrix factorization–free" method) derives from the well–known inverse iteration procedure in such a way that the associated system of linear equations is solved approximately by using a (multigrid) preconditioner.

A new convergence analysis for preconditioned inverse iteration is presented. The preconditioner is assumed to satisfy some bound for the spectral radius of the error propagation matrix resulting in a simple geometric setup. In this first part the case of poorest convergence depending on the choice of the preconditioner is analyzed. In the second part the dependence on all initial vectors having a fixed Rayleigh quotient is considered. The given theory provides sharp convergence estimates for the eigenvalue approximations showing that multigrid eigenvalue/vector computations can be done with comparable efficiency as known from multigrid methods for boundary value problems.

1. INTRODUCTION

The discretization of eigenvalue problems for partial differential operators leads to matrix eigenvalue problems having large dimensions in practice, fairly often more than 10^5 or 10^6. A finite element discretization, for instance, of an eigenvalue problem for a selfadjoint and coercive elliptic partial differential operator gives a generalized matrix eigenvalue problem of the form

$$Ax = \lambda M x,$$

where A, M are symmetric and positive definite matrices. A is called the discretization matrix and M is called the mass matrix. Typically, only a few of the smallest eigenvalues together with its eigenvectors are to be determined. In applications these eigenvalues are often the base frequencies of some vibrating mechanical structure, possibly of a turbine or an aircraft represented by finite element models.

The numerical treatment of these eigenvalue problems requires appropriate algorithms, since the matrices A and M are sparse with only a small, bounded number of nonzero elements per row. Therefore, these matrices are not stored explicitly, but only routines are provided to compute the matrix vector products Ax and Mx. Classical methods for the solution of the eigenvalue problem inasmuch they require any manipulation or factorization of A cannot be applied, since the computer storage for full matrices is not available. Hence, the QR method is not applicable. Moreover, the Lanczos method turns out to converge slowly since the condition number of A increases for decreasing mesh size h; for a 2D Laplacian on a uniform mesh the condition number behaves like h^{-2}. Finally, the Rayleigh

1991 *Mathematics Subject Classification.* Primary 65F15, 65N25; Secondary 65N30.

quotient method with its tempting cubic convergence in the eigenvalue approximations cannot be applied since the solution of equations within the shifted discretization matrix, which is then an indefinite matrix, is a critical step [29, 30, 33].

On the other hand, systems of linear equations within the discretization matrix can be solved efficiently by using multigrid or domain decomposition methods [2, 4, 31, 34]. The application of these methods can be represented by some approximate inverse, also called preconditioner, of the system matrix A. Therefore, in order to solve our eigenvalue problem we take up the well–known inverse iteration procedure and solve the associated system of linear equations in A approximately by using a preconditioner.

To introduce inverse iteration and for the following analysis we restrict the eigenvalue problem to the standard one, i.e. we set $M = I$, where I denotes the identity matrix. This assumption is nonrestrictive; the generalized eigenvalue problem is treated in [21]. Inverse iteration [5, 11, 12, 23] maps a given iterate x to the next iterate $\hat{x}$ by solving the system of linear equations

$$A\hat{x} = \lambda x, \tag{1.1}$$

with some subsequent normalization of $\hat{x}$. For our purposes we have slightly modified the standard representation of inverse iteration in a way that an additional scaling constant $\lambda = \lambda(x)$ appears on the right–hand side of (1.1). Therein $\lambda(x)$ denotes the Rayleigh quotient

$$\lambda(x) = \frac{(x, Ax)}{(x, x)} \tag{1.2}$$

of the actual nonzero iteration vector x. The constant λ in Equation (1.1) has no effect on the convergence properties of inverse iteration, but ensures stationarity ($\hat{x} = x$) in any eigenvector of A. It is well known that inverse iteration converges to the smallest eigenvalue λ_1 and to a corresponding eigenvector if the initial vector is not perpendicular to the invariant subspace of eigenvectors belonging to λ_1 [23].

To solve Equation (1.1) approximately we apply a symmetric and positive definite preconditioner B^{-1} for A which is assumed to satisfy

$$\|I - B^{-1}A\|_A \leq \gamma \tag{1.3}$$

for some constant γ with $0 \leq \gamma < 1$. Therein, $\|\cdot\|_A$ denotes the operator norm induced by A. The assumption (1.3) is typical for multigrid or domain decomposition preconditioners. (E.g. for A being the discretization of the Laplacian, a standard V–cycle with Jacobi smoothing leads to $\gamma \approx 0.2$.) The best preconditioners satisfy (1.3) with γ bounded away from 1 independently on the mesh size or the number of unknowns [34]. We note, in case of having a spectral equivalence

$$\gamma_1(x, Ax) \leq (x, Bx) \leq \gamma_2(x, Ax), \qquad \text{for all } x \neq 0, \quad \gamma_1, \gamma_2 > 0, \tag{1.4}$$

instead of (1.3), the following analysis is applicable to a scaled preconditioner [21].

The assumption (1.3) on the preconditioner B^{-1} expresses that the error propagation matrix $I - B^{-1}A$ is a reducer: In terms of the error propagation equation

$$x' - \lambda A^{-1}x = (I - B^{-1}A)(x - \lambda A^{-1}x), \tag{1.5}$$

$I - B^{-1}A$ being a reducer means that the initial error $x - \lambda A^{-1}x$, i.e. the difference of the vector x and the exact solution $\lambda A^{-1}x$ of (1.1), is reduced to the final error $x' - \lambda A^{-1}x$, where x' denotes the approximate solution of (1.1). In the case of the best possible preconditioner, i.e. $\gamma = 0$ or $B = A$, one has the maximal error reduction or, equivalently, in one step the result of inverse iteration $x' = \lambda A^{-1}x$.

We rewrite the error propagation equation in the form (containing no inverse of A)

$$x' = x - B^{-1}(Ax - \lambda x), \tag{1.6}$$

and call the iterative scheme *preconditioned inverse iteration* or abbreviated PINVIT. To iterate this scheme one has to provide routines computing matrix vector products with A and B^{-1}. These matrices are neither stored explicitly nor modified. For this reason PINVIT is a "matrix–free" method.

In this work we analyze the convergence behavior of preconditioned inverse iteration using the simple constraint (1.3) on the quality of the preconditioner. Our central task is to derive a sharp estimate Φ for the relative decrease of the Rayleigh quotient $\lambda(x')$ towards the next smaller eigenvalue λ_i in terms of eigenvalue approximations, as given by

$$\frac{\lambda(x') - \lambda_i}{\lambda - \lambda_i} \leq \Phi < 1. \tag{1.7}$$

Therefore it is assumed that λ_i and λ_{i+1} are the nearest eigenvalues of A enclosing λ, $\lambda_i < \lambda < \lambda_{i+1}$. To derive the sharp bound Φ one determines the supremum of the Rayleigh quotient $\lambda(x')$ with respect to the choice of the preconditioner as well as on the choice of x having the Rayleigh quotient λ. The important result is that Φ only depends on the two eigenvalues λ_i and λ_{i+1} enclosing λ as well as on γ and λ, i.e.

$$\Phi = \Phi(\lambda_i, \lambda_{i+1}, \gamma, \lambda).$$

This independence on all the other eigenvalues, and in particular the independence on the largest eigenvalue of A, qualifies PINVIT as an effective algorithm for grid eigenvalue problems, since the convergence estimate Φ can be bounded away from 1 and does not depend on the mesh size and hence the number of unknowns. Thus eigenvalue computations with preconditioned inverse iteration can be done with an efficiency as known from multigrid methods for boundary value problems [21]. By using the estimate on the eigenvalue approximations we can also determine a simple convergence estimate for the eigenvector approximations.

As expressed by Equation (1.7), the Rayleigh quotients of the iterates of PINVIT form a monotone decreasing sequence. For an initial vector x with $\lambda = \lambda(x) \in]\lambda_i, \lambda_{i+1}[$ the Rayleigh quotients of the iterates at least converge *linearly* down to λ_i by (1.7), but in the case of a faster decrease of the Rayleigh quotient they may jump from the interval $]\lambda_i, \lambda_{i+1}[$ to $[\lambda_1, \lambda_i[$. In principle, it cannot be said, when the Rayleigh quotients move from the interval $[\lambda_i, \lambda_{i+1}]$ to the next interval $[\lambda_{i-1}, \lambda_i]$ of smaller eigenvalues, since this depends on the actual choice of the preconditioner and on the (unknown) eigenvector expansion of the actual iterate. But in any case it is guaranteed that PINVIT converges to an eigenvector/value; usually as an effect of rounding errors to the smallest eigenvalue and a corresponding eigenvector. The convergence properties of PINVIT, its interpretation and how to define a *convergence rate* for PINVIT is discussed in detail in the introduction of Part II.

We do not claim to introduce a new or better converging eigensolver, but we hope that the analysis increases the understanding of what can be achieved with this form of preconditioning for the eigenproblem. We note that the iterative eigensolver analyzed here is in some sense the most simple one and that more refined preconditioning strategies for iterative eigenvalue solvers are known [27, 28]. Nevertheless, our theoretical analysis provides the basis for the convergence analysis of an analogous subspace iteration in [20], where sharp convergence estimates for the Ritz values belonging to the actual subspace are derived.

Furthermore, we do not discuss the question on how to construct or select an appropriate preconditioner for the PINVIT algorithm since this question is separated from our analysis by inequality (1.3) and the constant γ. There is no need to construct special preconditioners to solve our eigenvalue problem, since any (multigrid) preconditioner satisfying (1.3) or (1.4) will work. For a discussion of more practical questions arising while constructing an *adaptive* multigrid subspace eigensolver, see [21].

We emphasize that the iteration (1.6) is by no means new. It is known as a preconditioned gradient method. This naming derives from the fact that the gradient of the Rayleigh quotient is given by

$$\nabla\lambda(x) = \frac{2}{(x,x)}(Ax - \lambda(x)x). \tag{1.8}$$

Hence one expects that the Rayleigh quotient of the iterate x' with

$$x' = x - \omega(Ax - \lambda(x)x) \tag{1.9}$$

is decreased. The convergence depends on a proper choice of the scaling constant ω. A vast literature can be found on gradient methods for the eigenvalue problem, discussing different scaling strategies, convergence properties, adoption of the conjugate gradient method [1, 8, 10, 17, 18, 25, 32]. Nevertheless, gradient methods suffer from their poor convergence properties; for mesh eigenproblems the convergence rate converges to 1 if the mesh size h decreases to 0, [22].

Preconditioning of gradient methods (by premultiplying the residual by a preconditioner for A) leads to the iterative scheme (1.6) and results in substantially improved convergence properties, see the discussion above. Preconditioned gradient methods for the eigenvalue problem were first studied by Samokish [26] and later by Petryshyn [24]. Estimates on the convergence rate were given by Godunov et. al. [9] and D'yakonov et. al. [7, 6]. See Knyazev for a survey on preconditioned eigensolvers [13]. These preconditioned gradient methods have been generalized to a subspace iteration [19, 16, 3]. Applying the analysis of this work to the iterative subspace scheme of Bramble, Knyazev and Pasciak [3] one can remove some restrictive assumptions and can derive sharp estimates for the Ritz values [20].

This paper is organized as follows: In Section 2 we give a convenient representation of PINVIT and present its simple geometry. In Section 3 the multiple eigenvalue case is treated. In Section 4 a detailed analysis describing the points of suprema of the Rayleigh quotient with respect to the choice of the preconditioner is given. The points of suprema are characterized by a Lagrange multiplier ansatz based on constraints which derive from the geometric description of PINVIT. We obtain the surprising fact that these suprema are taken in points which can be represented by inverse iteration with a *positive* shift if applied to the given iterate. Finally Section 5 contains a mini–dimensional analysis of PINVIT which leads to sharp convergence estimates in $\mathbf{R}^2$.

In Part II we derive sharp convergence estimates for PINVIT. Therefore we vary not only the preconditioner but additionally the vector x whose Rayleigh quotient is assumed to have a fixed value. The analysis is based on the representation of the points of suprema gained in this part. Finally, by using predominantly geometric methods, we derive sharp convergence estimates for the Rayleigh quotient of the iterates. Additionally, we show that the acute angle between the actual iteration vector (i.e. the eigenvector approximation) and the invariant subspace to the smallest eigenvalue is not generally monotone decreasing in the course of the iteration. Nevertheless, the convergence of the eigenvector approximations results from the convergence of the eigenvalue approximations.

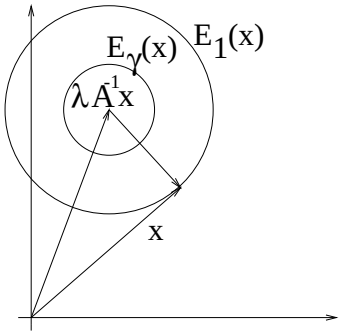

FIGURE 1. *The set $E_\gamma(x)$ with respect to the $\|\cdot\|_A$ norm.*

2. THE GEOMETRY OF PRECONDITIONED INVERSE ITERATION

Consider a symmetric positive definite matrix $A \in \mathbf{R}^{m\times m}$ with n different eigenvalues $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_n$ and assume the multiplicity of the i–th eigenvalue to be denoted by $m(i)$ so that $m = \sum_{i=1}^{n} m(i)$.

Furthermore let preconditioned inverse iteration be given by

$$x' = x - B^{-1}(Ax - \lambda x), \tag{2.1}$$

and assume that a symmetric and positive definite matrix B and a constant γ with $0 \le \gamma < 1$ are given so that

$$\|I - B^{-1}A\|_A \le \gamma \tag{2.2}$$

holds.

Applying preconditioned inverse iteration (2.1) to a given iterate x for all admissible preconditioners satisfying (2.2) gives rise to the definition of the set $E_\gamma(x)$ which contains all possible iterates

$$E_\gamma(x) := \{\lambda A^{-1}x + (I - B^{-1}A)(I - \lambda A^{-1})x;\, \|I - B^{-1}A\|_A \le \gamma\}. \tag{2.3}$$

In the following we analyze the extremal behavior of the Rayleigh quotient on the set $E_\gamma(x)$. The detailed analysis of this extremal behavior and its dependence on x finally leads to the required convergence estimates for PINVIT.

Figure 1 illustrates the set $E_\gamma(x)$ with respect to the vector norm induced by A. The next lemma provides some orthogonal decomposition and shows that the null vector is not contained in $E_\gamma(x)$.

Lemma 2.1. *For $x \in \mathbf{R}^m \setminus \{0\}$ holds*

(1) $(x, (I - \lambda A^{-1})x)_A = 0$,
(2) $\|\lambda A^{-1}x\|_A^2 = \|x\|_A^2 + \|(I - \lambda A^{-1})x\|_A^2$,
(3) $0 \notin E_\gamma(x)$ *for all* $\gamma \in [0, 1]$.

Therein, $\|\cdot\|_A$ and $(\cdot,\cdot)_A$ denote the norm and the inner product induced by A.

Proof. Properties (1) and (2) follow from

$$(x, (I - \lambda A^{-1})x)_A = (x, x)_A - \lambda(x)(x, A^{-1}x)_A = 0.$$

Using the triangle inequality, (2.2) and (2) give for nonzero x

$$\begin{aligned} \|x'\|_A &= \|\lambda A^{-1}x + (I - B^{-1}A)(I - \lambda A^{-1})x\|_A \\ &\geq \|\lambda A^{-1}x\|_A - \|(I - \lambda A^{-1})x\|_A \\ &= \left(\|\lambda A^{-1}x\|_A + \|(I - \lambda A^{-1})x\|_A\right)^{-1} \|x\|_A^2 > 0. \end{aligned}$$

□

In order to show that $E_\gamma(x)$ is a ball with respect to the $\|\cdot\|_A$–norm (whose center is $\lambda A^{-1}x$ and whose radius is $\|(I - \lambda A^{-1})x\|_A$) we construct a specific class of preconditioners built from Householder reflections.

Lemma 2.2. *Consider a Householder reflection $H = I - 2uu^T$ for $u \in \mathbf{R}^m$, $u^T u = 1$, and let $\hat{\gamma} \in [0, 1[$. Then*

$$\hat{B}^{-1} = A^{-1} + \hat{\gamma} A^{-1/2} H A^{-1/2} \tag{2.4}$$

is symmetric and positive definite and

$$\|I - \hat{B}^{-1}A\|_A = \hat{\gamma}.$$

Proof. Symmetry of $\hat{B}$ follows from the definition. For any nonzero $x \in \mathbf{R}^m$ and with $y = A^{-1/2}x$ follows

$$\begin{aligned} (x, \hat{B}^{-1}x) &= (x, A^{-1}x) + \hat{\gamma}(x, A^{-1/2}HA^{-1/2}x) = (y, y) + \hat{\gamma}(y, Hy) \\ &\geq (y, y) - \hat{\gamma}|y|\,|Hy| = (1 - \hat{\gamma})|y|^2 > 0 \end{aligned}$$

which shows that $\hat{B}$ is positive definite. Furthermore it holds that ($|\cdot|$ denotes the Euclidean norm)

$$\|(I - \hat{B}^{-1}A)x\|_A = \hat{\gamma}\|A^{-1/2}HA^{1/2}x\|_A = \hat{\gamma}|HA^{1/2}x| = \hat{\gamma}\|x\|_A.$$

□

Using these preconditioners $\hat{B}$ one obtains the required characterization of $E_\gamma(x)$.

Lemma 2.3. *$E_\gamma(x)$ is a ball with respect to the $\|\cdot\|_A$–norm with center $\lambda A^{-1}x$ and radius $\gamma\|(I - \lambda A^{-1})x\|_A$, i.e.*

$$E_\gamma(x) = \{\lambda A^{-1}x + y;\, y \in \mathbf{R}^m, \|y\|_A \leq \gamma\|(I - \lambda A^{-1})x\|_A\}.$$

Proof. By definition (2.3) obviously $E_\gamma(x)$ is a subset of the ball. To show the opposite inclusion consider a point $\lambda A^{-1}x + y$ in the ball. Then determine $\hat{\gamma}$ with $0 \leq \hat{\gamma} \leq \gamma$ from

$$\|y\|_A = \hat{\gamma}\|(I - \lambda A^{-1})x\|_A.$$

Moreover, a Householder reflection H can be determined which maps $\hat{\gamma}A^{1/2}(I - \lambda A^{-1})x$ to $-A^{1/2}y$ so that

$$-A^{1/2}y = \hat{\gamma}HA^{1/2}(I - \lambda A^{-1})x.$$

We conclude that $\lambda A^{-1}x + y \in E_\gamma(x)$ since

$$y = -\hat{\gamma}A^{-1/2}HA^{1/2}(I - \lambda A^{-1})x = (I - \hat{B}^{-1}A)(I - \lambda A^{-1})x,$$

using a preconditioner $\hat{B}$ as considered in Lemma 2. □

As a consequence of Lemma 2.3 preconditioners built from Householder reflections generate the complete ball $E_\gamma(x)$. We thus restrict the analysis of PINVIT to this type of preconditioners in order to simplify the iterative scheme and give its representation with respect to the basis of A–orthonormal eigenvectors of A in the next lemma.

Lemma 2.4. *Preconditioned inverse iteration (2.1) with the preconditioner (2.4) takes with respect to the A–orthonormal basis of eigenvectors of A the form*

$$c' = \lambda\Lambda^{-1}c - \hat{\gamma}(I - 2vv^T)(I - \lambda\Lambda^{-1})c, \tag{2.5}$$

where c and c' are the coefficient vectors within this basis of x and x', respectively. Moreover, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbf{R}^{m\times m}$, $\hat{\gamma} \leq \gamma$ and $v \in \mathbf{R}^m$, $|v| = 1$. The Rayleigh quotient of a nonzero $d \in \mathbf{R}^m$ with respect to this basis is given by

$$\lambda(d) = \frac{(d,d)}{(d,\Lambda^{-1}d)}. \tag{2.6}$$

Proof. Let X be the orthogonal matrix containing in the columns the eigenvectors of A so that $X^TAX = \Lambda$ and $X^TX = I$. Then for the coefficient vector c of x with respect to the basis of A–orthonormal eigenvectors of A holds

$$x = X\Lambda^{-1/2}c. \tag{2.7}$$

From (2.1) and for $B = \hat{B}$ by (2.4) we obtain

$$c' = c - \Lambda^{1/2}X^TB^{-1}X\Lambda^{1/2}(I - \lambda\Lambda^{-1})c = \lambda\Lambda^{-1}c - \hat{\gamma}X^THX(I - \lambda\Lambda^{-1})c. \tag{2.8}$$

Equations (2.5) follows since both H and X^THX are Householder reflections. Evaluating the Rayleigh quotient (1.2) of $X\Lambda^{-1/2}d$ results in (2.6). □

In the sequel, the convergence analysis is represented with respect to the c–basis introduced in Lemma (2.4). For the sake of convenience we define $E_\gamma(c)$ to be the c–basis representation (the basis introduced in Lemma 2.4) of $E_\gamma(x)$, i.e.

$$E_\gamma(c) := \{\Lambda^{1/2}X^Tz;\ z \in E_\gamma(x)\} = \{c' \text{ given by (2.5)}\}. \tag{2.9}$$

We finally note that the maximal Rayleigh quotient on $E_\gamma(c)$ does not depend on the sign of any component of c, since a change of the sign of the k–th component of c leads to a reflection of $E_\gamma(c)$ by a hyperplane orthogonal to the k–th unit vector through the origin. Furthermore, the Rayleigh quotient (2.6) is purely quadratic in the components of its argument so that any sign dependence vanishes.

Therefore, we restrict the convergence analysis to non–negative coefficient vectors c.

3. MULTIPLE EIGENVALUES

In this section we give a justification for restricting the further convergence analysis of PINVIT to matrices with only simple eigenvalues. Now let us write the diagonal matrix Λ, which contains the eigenvalues of A, in the form

$$\Lambda = \mathrm{diag}(\underbrace{\lambda_1, \ldots, \lambda_1}_{m(1)}, \ldots, \underbrace{\lambda_n, \ldots, \lambda_n}_{m(n)}) \in \mathbf{R}^{m\times m}.$$

In the same way let $d = (d_{1;1}, \ldots, d_{1;m(1)}, \ldots \quad, d_{n;1}, \ldots, d_{n;m(n)})^T$ for a coefficient vector $d \in \mathbf{R}^m$, where $d_{i;j}$ denotes the j–th component corresponding to the i–th eigenvalue of multiplicity $m(i)$. Now consider the mapping $P : \mathbf{R}^m \to \mathbf{R}^n$ which defines a problem of smaller dimension with simple eigenvalues by condensing components belonging to a multiple eigenvalue in a joint component.

$$(Pd)|_i = \bar{d}_i := (\sum_{j=1}^{m(i)} d_{i,j}^2)^{1/2}. \tag{3.1}$$

The Rayleigh quotient in the $\mathbf{R}^n$ with $\bar{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is denoted by

$$\bar{\lambda}(\bar{d}) = \frac{(\bar{d}, \bar{d})}{(\bar{d}, \bar{\Lambda}^{-1}\bar{d})}.$$

PINVIT for the reduced problem with $\bar{c} := P(c)$ reads

$$\bar{c}' = \bar{\lambda}(\bar{c})\bar{\Lambda}^{-1}\bar{c} - \hat{\gamma}\bar{H}(I - \bar{\lambda}(\bar{c})\bar{\Lambda}^{-1})\bar{c} \tag{3.2}$$

for arbitrary Householder reflections $\bar{H} \in \mathbf{R}^{n\times n}$. Note that (3.2) defines a ball $E_\gamma(\bar{c}) \subset \mathbf{R}^n$. The next lemma shows that the suprema in the case of simple eigenvalues dominate those of the multiple eigenvalue case. To make the convergence analysis of PINVIT complete, we show later in Section 3.3 of Part II that the sharp convergence estimates (as derived for the case of simple eigenvalues) are also sharp for matrices with eigenvalues of arbitary multiplicity.

Lemma 3.1. *Let $c \in \mathbf{R}^m$, then*

$$\sup \lambda(E_\gamma(c)) \leq \sup \bar{\lambda}(E_\gamma(\bar{c})).$$

Proof. From definition (3.1) we obtain by direct calculation

$$\lambda(d) = \bar{\lambda}(Pd) \qquad \text{for any } d \in \mathbf{R}^m \tag{3.3}$$

and thus $\lambda = \lambda(c) = \bar{\lambda}(Pc)$. Moreover, P maps the center of $E_\gamma(c)$ to the center of $E_\gamma(\bar{c})$, i.e. $P(\lambda\Lambda^{-1}c) = \lambda\bar{\Lambda}^{-1}\bar{c}$. Both balls have the same radius, since $|c - \lambda\Lambda^{-1}c| = |\bar{c} - \lambda\bar{\Lambda}^{-1}\bar{c}|$.

The mapping P preserves or reduces any distance, since for any $d, e \in \mathbf{R}^m$ (with $\bar{d} = Pd$ and $\bar{e} = Pe$) we have by using the Cauchy–Schwarz inequality

$$\begin{aligned}
|e - d|^2 &= \sum_{i=1}^{n}\sum_{j=1}^{m(i)}(e_{i;j} - d_{i;j})^2 \\
&\geq \sum_{i=1}^{n}\bar{e}_i^2 + \sum_{i=1}^{n}\bar{d}_i^2 - 2\sum_{i=1}^{n}\left((\sum_{j=1}^{m(i)}e_{i;j}^2)^{1/2}(\sum_{j=1}^{m(i)}d_{i;j}^2)^{1/2}\right) \\
&= |\bar{e} - \bar{d}|^2 = |Pe - Pd|^2.
\end{aligned}$$

The combination of all these geometric properties gives that $P(E_\gamma(c))$ is a subset of $E_\gamma(\bar{c})$. Hence

$$\sup \bar{\lambda}(P(E_\gamma(c))) \leq \sup \bar{\lambda}(E_\gamma(\bar{c})),$$

from which with (3.3) the proposition follows. □

4. Characterization of suprema of the Rayleigh quotient on $E_\gamma(c)$

By using the abbreviation (AC) we summarize three nonrestrictive assumptions on the vector $c \in \mathbf{R}^n$.

(AC) $\left\{\begin{array}{l} 1.\ |c|^2 = 1\ , \\ 2.\ c \text{ is not equal to any unit vector } e_i,\ i = 1, \ldots, n, \\ 3.\ c \text{ is componentwise nonnegative.} \end{array}\right.$

The first assumption is justified since PINVIT is homogeneous with respect to scaling. By the second assumption we exclude that PINVIT is stationary in an eigenvector. The third assumption is justified in Section 2.

4.1. **Localization of points of suprema in $E_\gamma(c)$.**

The gradient and the Hessian of the Rayleigh quotient are characterized by the following lemma.

Lemma 4.1. *For any nonzero $c \in \mathbf{R}^n$ the Rayleigh quotient (2.6) fulfills:*

(a) $\nabla\lambda(c) = \frac{2}{(c,\Lambda^{-1}c)}(I - \lambda\Lambda^{-1})c$.
(b) $\nabla\lambda(c) = 0$ *if and only if* $c = \theta e_i$, $1 \le i \le n$, *for some nonzero* $\theta \in \mathbf{R}$.
(c) *The Hessian $H(c)$ of (2.6) is given by*

$$H(c) = \frac{2}{(c,\Lambda^{-1}c)^2}\left[(I-\lambda\Lambda^{-1})(c,\Lambda^{-1}c) - 2(\Lambda^{-1}c)[(I-\lambda\Lambda^{-1})c]^T - 2[(I-\lambda\Lambda^{-1})c](\Lambda^{-1}c)^T\right]. \tag{4.1}$$

Proof. (a) and (c) follow from (2.6) by direct computation. Furthermore, all eigenvalues are simple so that (b) results. □

The next lemma shows that all points of suprema of the Rayleigh quotient on $E_\gamma(c)$, which represent the case of poorest convergence of PINVIT, are located on its surface or more precise on the surface of the circular cone $C_\gamma(c)$ enclosing $E_\gamma(c)$. The cone $C_\gamma(c)$ is defined by

$$C_\gamma(c) := \{\zeta d;\ d \in E_\gamma(c),\ \zeta > 0\}. \tag{4.2}$$

Since the Rayleigh quotient (2.6) is invariant with respect to nonzero scaling of its argument, the suprema with respect to $E_\gamma(c)$ and $C_\gamma(c)$ coincide.

Lemma 4.2. *Let (AC) be fulfilled and let $w \in \arg\sup\lambda(E_\gamma(c))$. Then $w \in \partial E_\gamma(c)$, i.e. the boundary of $E_\gamma(c)$.*

Proof. Let $w \in \arg\sup\lambda(E_\gamma(c))$ and assume w in the interior of $E_\gamma(c)$. Then $\nabla\lambda(w) = 0$ and thus $w = \theta e_i$ for a nonzero θ by Lemma 4.1. We first assume $i = n$ and derive a contradiction: The angle of opening ϕ of the circular cone $C_1(c)$ is given by

$$\cos\phi = \frac{(c,\lambda\Lambda^{-1}c)}{|c||\lambda\Lambda^{-1}c|} = \frac{1}{|\lambda\Lambda^{-1}c|}.$$

Furthermore, the acute angle χ enclosed by $\lambda\Lambda^{-1}c$ and $w = \theta e_n$ reads

$$\cos\chi = \frac{\lambda\lambda_n^{-1}c_n}{|\lambda\Lambda^{-1}c|}.$$

Since $|c| = 1$ by (AC), we have $\lambda\lambda_n^{-1}c_n < 1$ and so $\phi < \chi$. Hence, $\theta e_n \notin C_1(c)$ and thus $\theta e_n \notin E_\gamma(c)$. In the remaining cases, $1 \le i \le n-1$, the Hessian (4.1) in the stationary points θe_i is a diagonal matrix

$$H(\theta e_i) = \frac{2\lambda_i}{\theta^2}(I - \lambda_i\Lambda^{-1}),$$

which has at least one positive eigenvalue $\frac{2\lambda_i}{\theta^2}(1 - \frac{\lambda_i}{\lambda_n}) > 0$, so that $w = \theta e_i$ is not a point of a supremum. □

The fact that any point of a supremum is located on the boundary of $E_\gamma(c)$ leads to some orthogonal decomposition characterizing these points.

Theorem 4.3. *Let c satisfy (AC), $\gamma \in [0,1[$ and $w \in \arg\sup \lambda(E_\gamma(c))$. Then*

$$(w, w - \lambda\Lambda^{-1}c) = 0, \tag{4.3}$$

$$|w|^2 + |w - \lambda\Lambda^{-1}c|^2 = |\lambda\Lambda^{-1}c|^2, \tag{4.4}$$

$$|w - \lambda\Lambda^{-1}c| = \gamma|(I - \lambda\Lambda^{-1})c|, \tag{4.5}$$

$$|w| > |c|. \tag{4.6}$$

Proof. Assume $(w, w - \lambda\Lambda^{-1}c) \neq 0$, then κw (with $\kappa = \frac{(w,\lambda\Lambda^{-1}c)}{(w,w)} \neq 0$) is an element of the interior of $E_\gamma(c)$ because

$$|w - \lambda\Lambda^{-1}c|^2 - |\kappa w - \lambda\Lambda^{-1}c|^2 = \frac{1}{|w|^2}\left(|w|^2 - (w, \lambda\Lambda^{-1}c)\right)^2 > 0.$$

Moreover, $\lambda(\kappa w) = \lambda(w)$ holds, so that κw as a point of a supremum in the interior of $E_\gamma(c)$ contradicts Lemma 4.2. The orthogonal decomposition (b) is a direct consequence of (a). Equation (c) only expresses $w \in \partial E_\gamma(c)$, see Lemma 4.2. Finally,

$$|w|^2 = |\lambda\Lambda^{-1}c|^2 - \gamma^2|(I - \lambda\Lambda^{-1})c|^2 > |\lambda\Lambda^{-1}c|^2 - |(I - \lambda\Lambda^{-1})c|^2 = |c|^2.$$

□

4.2. **Characterization of suprema by the method of Lagrange multipliers.**

The next lemma explicitly describes the points of suprema by using the method of Lagrange multipliers.

Lemma 4.4. *Let c satisfy (AC), and assume $w \in \arg\sup \lambda(E_\gamma(c))$. Then there are constants $\mu,\ \nu \in \mathbf{R}$, so that*

$$2(\Lambda^{-1} + \mu + \nu)w = \nu\lambda\Lambda^{-1}c. \tag{4.7}$$

Proof. By (4.4) and (4.5) the left–hand side $|w|^2$ of

$$|w|^2 = |\lambda\Lambda^{-1}c|^2 - \gamma^2|(I - \lambda\Lambda^{-1})c|^2$$

has a fixed value for given c and γ. Hence, the function $\kappa(w) := (w, \Lambda^{-1}w)$ takes its extrema in the same arguments as the Rayleigh quotient $\lambda(w)$. The method of Lagrange multipliers applied to $\kappa(w)$ with respect to the constraints (4.3) and (4.4) leads to a Lagrange function $L = L(w, \mu, \nu)$ with multipliers μ and ν in the form

$$L = (w, \Lambda^{-1}w) + \mu\left(|w|^2 + \gamma^2|(I - \lambda\Lambda^{-1})c|^2 - |\lambda\Lambda^{-1}c|^2\right) + \nu(w, w - \lambda\Lambda^{-1}c).$$

The gradient of L with respect to w reads

$$\nabla L = 2(\Lambda^{-1} + \mu + \nu)w - \nu\lambda\Lambda^{-1}c.$$

From $\nabla L = 0$ the assertion follows. □

The following analysis distinguishes the cases $\nu \neq 0$ and $\nu = 0$. Next we treat $\nu = 0$.

Lemma 4.5. *Let c satisfy (AC) and let $w \in \arg\sup \lambda(E_\gamma(c))$. Assuming $\nu = 0$ in (4.7) any w has the form*

$$w = \lambda\lambda_k^{-1}c_k e_k,$$

for some k with $1 < k < n$ and nonzero c_k. (e_k denotes the k–th unit vector.)

Proof. For $\nu = 0$ from (4.7) we have

$$(\Lambda^{-1} + \mu)w = 0.$$

By (4.6) the vector w has at least one nonzero component w_k for $1 \le k \le n$. Hence, $\mu = -1/\lambda_k$. All eigenvalues λ_i are simple, so that $w = \theta e_k$ for nonzero θ. From (4.3) we obtain $\theta = \lambda\lambda_k^{-1}c_k$. We also have $c_k \neq 0$ since otherwise $w = 0$.

The cases $k = 1$ and $k = n$ are impossible: For $k = 1$ a point of a supremum in $w = \lambda\lambda_1^{-1}c_1e_1$ contradicts $\lambda(\lambda\Lambda^{-1}c) > \lambda_1 = \lambda(w)$. Furthermore, the proof of Lemma 4.2 excludes for $k = n$ a point of a supremum in $w = \lambda\lambda_k^{-1}c_ne_n$. □

Remark 4.6. Assuming $\nu = 0$ in Lemma 4.5 leads to a maximum of $n - 2$ candidates for points of suprema of the Rayleigh quotient on $E_\gamma(c)$. In the sequel we assume $\nu \neq 0$ and derive a continuum of points of extrema depending on a real parameter. Later in Appendix A of Part II we show that only these points are suprema and not the candidates obtained in Lemma 4.5.

Lemma 4.7. *Let c satisfy (AC), $w \in \arg\sup\lambda(E_\gamma(c))$ and assume $\nu \neq 0$. Then for any positive component $c_k > 0$ of the nonnegative vector c holds*

$$w_k = \frac{\lambda\nu}{2(1 + \lambda_k(\mu + \nu))}c_k > 0. \tag{4.8}$$

Furthermore, if $c_k = 0$ then $w_k = 0$ for $k = 1, \dots, n-1$.

Proof. If $\nu c_k \neq 0$, then $\lambda_k^{-1} + \mu + \nu$ and w_k are nonzero by (4.7) from which the form of w_k follows. If $w_k < 0$, then define $\bar{w}$ to be equal to w but with a positive sign of the k–th component. A comparison of the distances of w and $\bar{w}$ to the center $\lambda\Lambda^{-1}c$ of the ball $E_\gamma(c)$ shows

$$|w - \lambda\Lambda^{-1}c|^2 - |\bar{w} - \lambda\Lambda^{-1}c|^2 = -4w_k\lambda\lambda_k^{-1}c_k > 0,$$

so that $\bar{w}$ is an element of the interior of $E_\gamma(c)$. Moreover, $\lambda(w) = \lambda(\bar{w})$, which contradicts Lemma 4.2 since all points of absolute extrema are located on the boundary of $E_\gamma(c)$. Hence $w_k > 0$.

Next assume $c_k,\ c_{k'} = 0$ and $w_k,\ w_{k'} \neq 0$. Then (4.7) implies

$$(\lambda_k^{-1} + \mu + \nu)w_k = 0 = (\lambda_{k'}^{-1} + \mu + \nu)w_{k'},$$

so that $\lambda_k = \lambda_{k'}$, or equivalently $k = k'$. Hence there is at most one component for which $c_k = 0$ and $w_k \neq 0$.

Now l denote the smallest index so that $c_l > 0$ and let l' the largest index with $c_{l'} > 0$. We assume $c_k = 0$ and $w_k \neq 0$ for $l < k < l'$. From (4.7) we deduce $\mu + \nu = -1/\lambda_k$ and thus obtain for w_l and $w_{l'}$

$$w_l = \frac{\nu\lambda_k\lambda}{\lambda_k - \lambda_l}\,\frac{c_l}{2}, \qquad w_{l'} = \frac{\nu\lambda_k\lambda}{\lambda_k - \lambda_{l'}}\,\frac{c_{l'}}{2}.$$

Since c_l, $c_{l'}$, w_l and $w_{l'}$ are positive and $\lambda_l < \lambda_k < \lambda_{l'}$ one obtains

$$\nu = \frac{w_l}{c_l}\,\frac{2(\lambda_k - \lambda_l)}{\lambda_k\lambda} > 0, \qquad \nu = \frac{w_{l'}}{c_{l'}}\,\frac{2(\lambda_k - \lambda_{l'})}{\lambda_k\lambda} < 0,$$

which contradicts $\nu \neq 0$. Hence $w_k = 0$.

Now consider the case $c_m = 0$ and $w_m \neq 0$ with $l' < m < n$. Define $\bar{w}$ to be equal to w with exception of the components with indexes m and n which have changed their places. Since $c_m = c_n = 0$ one has

$$|\lambda\Lambda^{-1}c - w| = |\lambda\Lambda^{-1}c - \bar{w}|,$$

and thus $\bar{w} \in E_\gamma(c)$ holds, but due to $\lambda_m < \lambda_n$ we have $\lambda(\bar{w}) > \lambda(w)$, which contradicts $w \in \arg\sup \lambda(E_\gamma(c))$.

In the remaining case $m < l$, define $\bar{w}$ to be equal to w except for the m–th component which is set equal to zero. Since $c_1 = \ldots = c_m = 0$, and thus $\lambda(\lambda\Lambda^{-1}c) > \lambda_m$, we conclude $\lambda(w) = \sup \lambda(E_\gamma(c)) > \lambda_m$. Hence, from $\lambda_m < \frac{(w,w)}{(w,\Lambda^{-1}w)}$ we obtain

$$[(w,w) - w_m^2](w,\Lambda^{-1}w) > [(w,\Lambda^{-1}w) - w_m^2/\lambda_m](w,w)$$

and thus $\lambda(\bar{w}) > \lambda(w)$. Additionally, $\bar{w} \in E_\gamma(c)$ since

$$|w - \lambda\Lambda^{-1}c|^2 - |\bar{w} - \lambda\Lambda^{-1}c|^2 = w_m^2 > 0$$

which contradicts $w \in \arg\sup \lambda(E_\gamma(c))$. □

In the next theorem we show that any point of a supremum has the very simple representation (4.9). So we obtain the somewhat surprising result that any point of a supremum w can be represented by application of inverse iteration with a shift to c. A similar analysis shows that such a result doesn't hold in general for points of infima.

Theorem 4.8. *On the assumptions of Lemma 4.7 any $w \in \arg\sup \lambda(E_\gamma(c))$ can be represented as resulting from inverse iteration with a shift, i.e. there are real constants α, $\beta \in \mathbf{R}$ such that*

$$w = \beta(\alpha + \Lambda)^{-1}c. \tag{4.9}$$

Proof. If $\gamma = 0$ then $w = \lambda\Lambda^{-1}c$, so that $\alpha = 0$ and $\beta = \lambda$. If $\gamma > 0$ then due to Lemma 4.7 representation (4.9) may only be violated assuming $c_n = 0$ together with $w_n \neq 0$. From (4.7) we have $\mu + \nu = -1/\lambda_n$. Hence the remaining components $w_1, \ldots, w_{n-1}$ read

$$w_i = \frac{\nu\lambda c_i}{2(1 - \lambda_i\lambda_n^{-1})}. \tag{4.10}$$

Inserting (4.10) in (4.3) we obtain for $|w|^2$

$$|w|^2 = \sum_{i\neq n} \frac{\nu\lambda^2 c_i^2}{2\lambda_i(1 - \lambda_i\lambda_n^{-1})} > 0.$$

We have $\nu = \frac{|w|^2}{\omega}$ with $\omega := \sum_{i\neq n} \frac{\lambda^2 c_i^2}{2\lambda_i(1-\lambda_i\lambda_n^{-1})}$. Elimination of ν in (4.10) results in

$$w_i = \frac{|w|^2\lambda c_i}{2\omega(1 - \lambda_i\lambda_n^{-1})}.$$

Then for w_n holds

$$w_n^2 = |w|^2 - \sum_{i\neq n} w_i^2 = |w|^2 \left(1 - \frac{|w|^2}{\omega^2} \sum_{i\neq n} \frac{\lambda^2 c_i^2}{4(1 - \lambda_i\lambda_n^{-1})^2}\right).$$

We show next that

$$\omega^2 \leq |w|^2 \sum_{i\neq n} \frac{\lambda^2 c_i^2}{4(1 - \lambda_i\lambda_n^{-1})^2}, \tag{4.11}$$

which implies $w_n^2 < 0$ in contradiction to $w_n^2 > 0$. Using (4.6) we see that from

$$\lambda^2 \left(\sum_{i\neq n} \frac{c_i^2}{\lambda_i(1 - \lambda_i\lambda_n^{-1})}\right)^2 \leq \left(\sum_{i\neq n} c_i^2\right) \sum_{i\neq n} \frac{c_i^2}{(1 - \lambda_i\lambda_n^{-1})^2},$$

inequality (4.11) follows. The last inequality is equivalent to

$$\sum_{i\neq n} c_i^2 \left(\sum_{i\neq n} \frac{c_i^2}{\lambda_i(\lambda_n-\lambda_i)} \right)^2 \leq \left(\sum_{i\neq n} \frac{c_i^2}{\lambda_i} \right)^2 \sum_{i\neq n} \frac{c_i^2}{(\lambda_n-\lambda_i)^2}. \tag{4.12}$$

In Appendix A Lemma A.2 proves inequality (4.12) for $k = n-1$ and $\tau = \lambda_n$. □

4.3. **Continuous curve of points of suprema.**

In the previous section points of suprema are shown to be of the form $w = \beta(\alpha+\Lambda)^{-1}c$ for real constants α and β. But so far the constants α and β are unknown. In this section we determine the constant β and show that there is a unique α for each $\gamma \in [0,1[$.

Lemma 4.9. *Let $c \in \mathbf{R}^n$ (with $n \geq 2$) satisfy (AC). Then the function*

$$\rho : \ [0,\infty[\to \mathbf{R} : \alpha \mapsto \lambda((\alpha+\Lambda)^{-1}c)$$

is strictly monotone increasing in α. Therein $\lambda(\cdot)$ denotes the Rayleigh quotient (2.6). Moreover, $\rho([0,\infty[) = [\lambda(\Lambda^{-1}c), \lambda(c)[$.

Proof. The diagonal matrix $(\alpha+\Lambda)$ is invertible for $\alpha \geq 0$. Hence consider $0 \leq \alpha_1 < \alpha_2$ be given and define $w^{(1)} := (\alpha_1+\Lambda)^{-1}c$ and $w^{(2)} := (\alpha_2+\Lambda)^{-1}c$. Then we have for $i = 1, \ldots, n$

$$w_i^{(1)} = \frac{\alpha_2+\lambda_i}{\alpha_1+\lambda_i} w_i^{(2)},$$

wherein the sequence of positive coefficients $\frac{\alpha_2+\lambda_1}{\alpha_1+\lambda_1}, \ldots, \frac{\alpha_2+\lambda_n}{\alpha_1+\lambda_n}$ is strictly monotone decreasing. Hence, due to Lemma A.1 the function ρ is strictly monotone increasing. Furthermore,

$$\rho(0) = \lambda(\Lambda^{-1}c) \quad \text{and} \quad \lim_{\alpha\to\infty} \lambda((\alpha+\Lambda)^{-1}c) = \lambda(c).$$

□

By using Lemma (4.9) we see in the next theorem that the points of suprema represent a continuous curve as a function of γ. The curve connects the center $\lambda\Lambda^{-1}c$ of $E_\gamma(c)$ for $\gamma = 0$ and the vector c for $\gamma = 1$.

Theorem 4.10. *Let c satisfy (AC). Then for each $\gamma \in [0,1[$ there are unique $\alpha \geq 0$ and $\beta > 0$ with*

$$\beta = \beta(\alpha) = \frac{(\lambda\Lambda^{-1}c, (\alpha+\Lambda)^{-1}c)}{((\alpha+\Lambda)^{-1}c, (\alpha+\Lambda)^{-1}c)},$$

so that $w = \beta(\alpha+\Lambda)^{-1}c \in \arg\sup \lambda(E_\gamma(c))$. Furthermore, this w is the unique point of a supremum on $E_\gamma(c)$.

Proof. For $w = \beta(\alpha+\Lambda)^{-1}c$ we have $\beta/(\alpha+\lambda_i) > 0$ for any nonzero component c_i by Lemma 4.7. If $\beta < 0$, then $\alpha < -\lambda_l$ (where l is the largest index so that $c_l > 0$) and the sequence $\frac{\beta}{\alpha+\lambda_i}$ only for indexes i with $c_i > 0$ is strictly monotone increasing. Hence from Lemma A.1 one obtains $\lambda(w) > \lambda(c)$. Such a result contradicts the convergence estimates of D'yakonov and Orekhov [7], since adapting that convergence analysis to the given assumption (1.3) and removing the scaling constant shows that the Rayleigh quotient never increases while applying PINVIT to c. Thus $\beta > 0$ and $\alpha > -\lambda_{\bar{l}}$, where $\bar{l}$ is the smallest index so that $c_{\bar{l}} \neq 0$. Since

$$\rho :]-\lambda_{\bar{l}}, \infty[\to \mathbf{R} : \alpha \mapsto \lambda((\alpha+\Lambda)^{-1}c)$$

is strictly monotone increasing in α (confer Lemma 4.9) and since for $\alpha = 0$ we have $\lambda(\beta\Lambda^{-1}c) = \lambda(\lambda\Lambda^{-1}c) = \lambda(E_0(c))$ we conclude that only nonnegative α may represent points of a suprema. Furthermore, the form of $\beta > 0$ directly follows from (4.3). Uniqueness of the point of a supremum follows from the fact that $\rho(\alpha)$ for positive α is strictly monotone increasing. □

4.4. **Reduction to a lower dimensional positive problem.**

As a result of Theorem 4.10 the case of poorest convergence of PINVIT can be represented by inverse iteration with a positive shift if applied to the given iterate. Hence zero components of c remain to be zero components of w. For this reason the convergence analysis of PINVIT can be restricted to the nonzero part of c, i.e. to the contributing eigenfunctions. The next lemma formally describes the reduction of the dimension and provides the basis for the convergence analysis of PINVIT in Part II. There we show that the Rayleigh quotient of the new iterate of PINVIT, under all $c \in \mathbf{R}^n$ with a fixed Rayleigh quotient, takes its maximum in a vector with only two nonzero components. Applying the mini–dimensional analysis given in the next section one finally obtains sharp convergence estimates.

For a given nonnegative $c \in \mathbf{R}^n$ let S_c be the operator which reduces the dimension of a vector $v \in \mathbf{R}^n$ by eliminating all components of v which are zero components of c. In the same way S_c is applied to the diagonal matrix Λ leading to a diagonal matrix of lower dimension. If, for example, $c = (c_1, 0, 0, c_4)^T$, $c_1, c_4 \neq 0$ then $S_c(v_1, \dots, v_4)^T = (v_1, v_4)^T$. The next lemma describes the geometry of $E_\gamma(S_c c)$.

Lemma 4.11. *Let $c \in \mathbf{R}^n \setminus \{0\}$ be a nonnegative vector and $d := S_c c$, $\Lambda_d := S_c \Lambda$. Then*

$$
\begin{aligned}
&(a) \qquad \lambda = \lambda(c) = \frac{(c,c)}{(c,\Lambda^{-1}c)} = \frac{(d,d)}{(d,\Lambda_d^{-1}d)},\\
&(b) \qquad S_c(\lambda\Lambda^{-1}c) = \lambda\Lambda_d^{-1}S_c(c) = \lambda\Lambda_d^{-1}d,\\
&(c) \qquad S_c((I-\lambda\Lambda^{-1})c) = (I-\lambda\Lambda_d^{-1})d,\\
&(d) \qquad |(I-\lambda\Lambda^{-1})c| = |(I-\lambda\Lambda_d^{-1})d|,\\
&(e) \qquad S_c(E_\gamma(c)) = E_\gamma(d).
\end{aligned}
$$

Hence the suprema of the Rayleigh quotient on $E_\gamma(c)$ and on $E_\gamma(S_c c)$ coincide.

Proof. Properties (a)–(e) follow from the definition of S_c. The rest follows from Theorem 4.10. □

5. MINI–DIMENSIONAL CONVERGENCE ANALYSIS OF PINVIT

The objective of this section is to derive a sharp convergence estimate for preconditioned inverse iteration in the case of the smallest nontrivial dimension of the eigenvalue problem, that is in the $\mathbf{R}^2$. The following "mini–dimensional analysis" is a first step towards a complete analysis of PINVIT. In Part II the convergence estimate given here turns out to be fundamental for the analysis in the $\mathbf{R}^n$. The concept of a mini–dimensional analysis is in some sense typical of the analysis of iterative eigensolvers. It is well–known that the convergence rate of the power method (or inverse iteration) is determined by the two largest (or by the two smallest) eigenvalues of the given matrix. A simple proof shows that the convergence rate estimate takes its maximal value in exactly those vectors which belong to that extremal eigenvalues. For the steepest ascent method (without preconditioning) to determine the maximal eigenvalue of a given matrix, Knyazev and Skorokhodov [15]

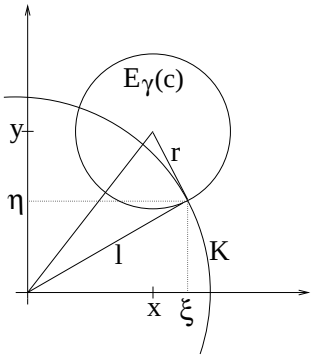

FIGURE 2. *Geometric setup.*

have also used a mini–dimensional technique to determine the convergence rate; for mini–dimensional analysis of a steepest descent method for systems of linear equations see [14].

Theorem 5.1. *Let $c \in \mathbf{R}^2$ with $\lambda_1 < \lambda = \lambda(c) < \lambda_2$ and $|c| = 1$. Let c' be defined by (2.5) as the result of preconditioned inverse iteration with a preconditioner fulfilling (2.2) for some $\gamma \in [0, 1[$. Then*

$$\lambda(c') \leq \lambda_{12}(\lambda, \gamma) \tag{5.1}$$

with

$$\lambda_{12}(\lambda, \gamma) = \frac{\lambda_1 \lambda_2}{\lambda_2 - \frac{\lambda_2 - \lambda_1}{1+m^2}}. \tag{5.2}$$

Therein m is the slope of that straight line through the origin and through $E_\gamma(c)$ which maximizes the Rayleigh quotient. It holds

$$m = \frac{yl - rx}{xl + ry} \tag{5.3}$$

where x, y, r and l are given by Equations (5.8)–(5.10).

One explicitly obtains λ_{12} as a function of λ, γ, λ_1 and λ_2 in the form

$$\begin{aligned} \lambda_{12}(\lambda, \gamma) \quad := \quad & \lambda\lambda_1\lambda_2(\lambda_1 + \lambda_2 - \lambda)^2 / \\ & \big(\gamma^2(\lambda_2 - \lambda)(\lambda - \lambda_1)(\lambda\lambda_2 + \lambda\lambda_1 - \lambda_1^2 - \lambda_2^2) \\ & -2\gamma\sqrt{\lambda_1\lambda_2}(\lambda - \lambda_1)(\lambda_2 - \lambda) \\ & \qquad \sqrt{\lambda_1\lambda_2 + (1-\gamma^2)(\lambda - \lambda_1)(\lambda_2 - \lambda)} \\ & -\lambda(\lambda_1 + \lambda_2 - \lambda)(\lambda\lambda_2 + \lambda\lambda_1 - \lambda_1^2 - \lambda_1\lambda_2 - \lambda_2^2)\big)\,. \end{aligned} \tag{5.4}$$

The estimate is sharp in a way that a preconditioner fulfilling (2.2) can be constructed such that $\lambda(c') = \lambda_{12}(\lambda, \gamma)$.

Proof. Due to Theorem (4.3) our task is to determine the unique point of intersection of $E_\gamma(c)$ with a straight line through the origin which is tangential to $E_\gamma(c)$ and maximizes the Rayleigh quotient. The geometric setup of the problem is shown in Figure 2. Therefore, we first construct the points of intersection of the circle $E_\gamma(c)$ with radius $r := \gamma|(I - \lambda\Lambda^{-1})c|$ with a second circle K of radius $l := \sqrt{x^2 + y^2 - r^2}$ centered at the

origin; therein the center of $E_\gamma(c)$ is given by $(y,x)^T = \lambda\Lambda^{-1}c$. The point of intersection maximizing the Rayleigh quotient on $E_\gamma(c)$ has the form

$$(\eta,\xi) = (\sqrt{l^2-\xi^2},\ \frac{xl^2+ryl}{x^2+y^2}). \tag{5.5}$$

Thus the Rayleigh quotient (2.6) of $(\eta,\xi)^T$ reads

$$\begin{aligned}\lambda_{12}(\lambda,\gamma) &= \lambda((\eta,\xi)^T) = \frac{\eta^2+\xi^2}{\eta^2/\lambda_1+\xi^2/\lambda_2} \\ &= \frac{\lambda_1\lambda_2(x^2+y^2)^2}{\lambda_2(x^2+y^2)^2+(\lambda_1-\lambda_2)(lx+yr)^2},\end{aligned} \tag{5.6}$$

from which we obtain (5.2) and (5.3).

The components of the positive vector $c\in\mathbf{R}^2$ are determined by $|c|=1$ and $\lambda(c)=\lambda$. Hence,

$$c_1 = \left(\frac{\lambda_1(\lambda_2-\lambda)}{\lambda(\lambda_2-\lambda_1)}\right)^{1/2}, \qquad c_2 = \left(\frac{\lambda_2(\lambda-\lambda_1)}{\lambda(\lambda_2-\lambda_1)}\right)^{1/2}. \tag{5.7}$$

For the center of $E_\gamma(c)$ one obtains $(y,x)^T = \lambda\Lambda^{-1}(c_1,c_2)^T$ or

$$x = \sqrt{\frac{\lambda(\lambda-\lambda_1)}{\lambda_2(\lambda_2-\lambda_1)}}, \qquad y = \sqrt{\frac{\lambda(\lambda_2-\lambda)}{\lambda_1(\lambda_2-\lambda_1)}}. \tag{5.8}$$

Thus for the radius r holds

$$r = \gamma|(I-\lambda\Lambda^{-1})c| = \gamma\sqrt{\frac{(\lambda-\lambda_1)(\lambda_2-\lambda)}{\lambda_1\lambda_2}}. \tag{5.9}$$

Finally, we have

$$l = \sqrt{\frac{\gamma^2(\lambda_1-\lambda)(\lambda_2-\lambda)+\lambda(\lambda_1+\lambda_2-\lambda)}{\lambda_1\lambda_2}}. \tag{5.10}$$

Inserting (5.8), (5.9) and (5.10) in (5.6) we obtain after some tedious but elementary simplifications $\lambda_{12}(\lambda,\gamma)$ in the form (5.4). Finally, by Lemma 2.2 a Householder reflection exists, so that c is mapped in the point of intersection $c' = (\eta,\xi)^T$ so that $\lambda(c') = \lambda_{12}(\lambda,\gamma)$. □

We note that with respect to the initial basis the theorem says that for $x\in\mathbf{R}^2$ (with $\lambda_1<\lambda=\lambda(x)<\lambda_2$) and a preconditioner B^{-1} fulfilling (2.2) for the Rayleigh quotient of the iterate x' by (2.1) the sharp estimate

$$\lambda(x') \le \lambda_{12}(\lambda,\gamma)$$

holds.

The function λ_{12} has two representations: In Equation (5.2) the slope m is the decisive factor. We have $\lambda_{12}=\lambda_2$ for $m=0$ and $\lambda_{12}\to\lambda_1$ as $m\to\infty$. To understand the dependence of m on γ one observes that $m=y/x$ for $\gamma=0$, which is the result of inverse iteration, and that $m=c_1/c_2$ for $\gamma=1$, which corresponds to stationarity of PINVIT. For $\gamma\in]0,1[$ the slope m depends on γ as described by Equations (5.3), (5.8)–(5.10). The square roots in r and l are responsible for the somewhat unreadable representation of λ_{12} by Equation (5.4), which results from (5.6) by inserting the geometric quantities and performing then extensive and tedious simplifications. It may be seen as a drawback of

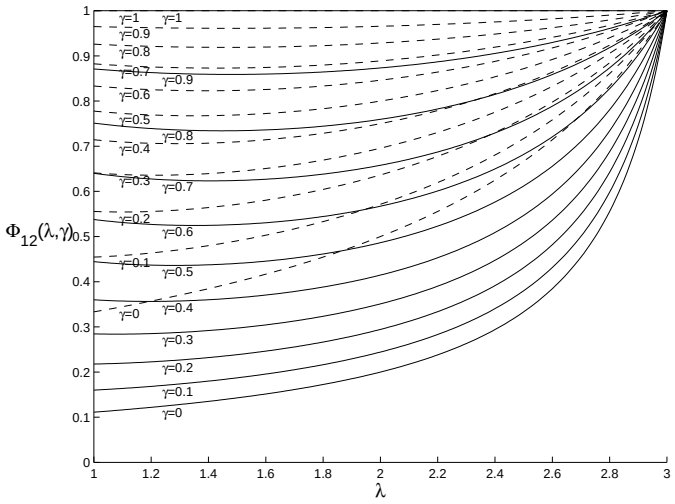

FIGURE 3. *Comparison of convergence estimates. Abscissa:* $\lambda \in [\lambda_1, \lambda_2] = [1, 3]$. *Solid lines: Optimal convergence estimate* $\Phi_{12}(\lambda, \gamma)$ *defined by (5.12). Broken lines: Estimate* $\hat{\Phi}(\lambda, \gamma)$ *by Equation (5.11).*

this analysis that (5.4) is a lengthy formula, since it is not easy to see by direct calculation that $\lambda_{12}(\lambda, \gamma) < \lambda$, which implies convergence of PINVIT.

Therefore, we conclude this section with a comparison of the classical convergence estimate by D'yakonov and Orekhov [7] and the estimate (5.4) which turns out as a significant improvement. By using the assumption (1.3) on the preconditioner and with a scaling constant $\omega = \frac{1}{1+\gamma}$ the analysis in [7] leads to the following estimate for the relative decrease of $\lambda(x')$ to λ_1

$$\frac{\lambda(x') - \lambda_1}{\lambda - \lambda_1} \leq \frac{1 - \frac{1-\gamma}{1+\gamma}\frac{\lambda_2 - \lambda}{\lambda_2}}{1 + \frac{1-\gamma}{1+\gamma}\frac{(\lambda - \lambda_1)(\lambda_2 - \lambda)}{\lambda_1 \lambda_2}} =: \hat{\Phi}(\gamma, \lambda). \tag{5.11}$$

Now we compare the convergence estimate $\hat{\Phi}(\lambda, \gamma)$ and the optimal estimate $\Phi_{12}(\lambda, \gamma)$

$$\Phi_{12}(\lambda, \gamma) := \frac{\lambda_{12}(\lambda, \gamma) - \lambda_1}{\lambda - \lambda_1} \tag{5.12}$$

with $\lambda_{12}(\lambda, \gamma)$ derived in Theorem 5.1. As a concrete example we take $\lambda_1 = 1$ and $\lambda_2 = 3$. In Figure 3 for $\gamma = \frac{k}{10}$, $k = 0, \ldots, 10$, the optimal estimate $\Phi_{12}(\lambda, \gamma)$ is charted by solid lines while $\hat{\Phi}(\lambda, \gamma)$ is represented by broken lines. Anticipating the results of Part II we note that $\Phi_{12}(\lambda, \gamma)$ as derived by the mini–dimensional analysis remains to be the optimal estimate in the $\mathbf{R}^n$. For this reason we make a comparison with the estimate in [7] and not with the more recent estimate (6.4) in [13]. The latter estimate does not only depend on the two nearest eigenvalues enclosing the Rayleigh quotient of the given iterate but also on the largest eigenvalue.

For $\gamma = 0$ the estimate $\Phi_{12}(\lambda, 0)$ corresponds to inverse iteration and derives from (5.12) and (5.4) for $\gamma = 0$.

$$\Phi_{12}(\lambda, 0) = \frac{\lambda(\lambda A^{-1}x) - \lambda_1}{\lambda(x) - \lambda_1} = \frac{\lambda_1^2}{\lambda_1^2 + (\lambda_2 - \lambda)(\lambda_1 + \lambda_2)} < 1. \tag{5.13}$$

In the limiting case $\gamma = 1$ the convergence estimate $\Phi_{12}(\lambda, 1)$ equals 1. Then PINVIT is stationary. Let us compare the convergence estimates for two situations, explicitly. If $\lambda = 2.0$ and $\gamma = 0.1$ one obtains $\hat{\Phi} \approx 0.571$ and $\Phi_{12} \approx 0.244$, while for $\lambda = 1.2$ and $\gamma = 0.2$ holds $\hat{\Phi} \approx 0.556$ and $\Phi_{12} \approx 0.223$.

6. Conclusion

Application of PINVIT to a given initial vector with a preconditioner satisfying the simple constraint (1.3) leads to a ball of iterates, where the center is defined by the result of inverse iteration. The Rayleigh quotient on this ball takes its supremum in a vector which can be represented as resulting from application of INVIT with a positive shift to the initial vector. For the smallest nontrivial dimension a sharp convergence estimate for PINVIT has been given.

In Part II we analyze the dependence of these suprema on all those initial vectors whose Rayleigh quotient has a fixed value. We finally derive sharp convergence estimates for PINVIT by applying the results of the mini–dimensional analysis given here.

Appendix A. Inequalities on weighted means of eigenvalues

We give two auxiliary lemmas used in Theorem 4.8 and in Section 4.3. The first lemma investigates the effect of a monotonous weighting function on the Rayleigh quotient.

Lemma A.1. *Let $c \in \mathbf{R}^n$ and let the Rayleigh quotient $\lambda(\cdot)$ be given by (2.6). Moreover, define $d \in \mathbf{R}^n$ by $d_i := a_i c_i$ for $i = 1, \ldots, n$ with a monotone increasing sequence of positive numbers $0 < a_1 \leq a_2 \leq \ldots \leq a_n$. Then for the Rayleigh quotients of c and d holds that*

$$\lambda(c) \leq \lambda(d).$$

Furthermore, if there are nonzero c_i, c_j for $i < j$ with $a_i < a_j$, then we even have $\lambda(c) < \lambda(d)$. Analogously, if the a_i are monotone decreasing the Rayleigh quotient decreases.

Proof. If $\lambda(c) = \lambda_n$ then $c = \theta e_n$ ($\theta \neq 0$) and $\lambda(c) = \lambda(d)$. Thus assume $\lambda(c) < \lambda_n$. Hence there is a unique m, so that $\lambda_m \leq \lambda(c) < \lambda_{m+1}$. Writing the Rayleigh quotient of d in the form

$$\lambda(d) = \frac{\sum_{i=1}^n d_i^2}{\sum_{i=1}^n d_i^2/\lambda_i} = \frac{\sum_{i<m} \frac{a_i^2}{a_m^2} c_i^2 + c_m^2 + \sum_{i>m} \frac{a_i^2}{a_m^2} c_i^2}{\sum_{i<m} \frac{a_i^2}{a_m^2} c_i^2/\lambda_i + c_m^2/\lambda_m + \sum_{i>m} \frac{a_i^2}{a_m^2} c_i^2/\lambda_i},$$

we have $\left(\frac{a_i}{a_m}\right)^2 \leq 1$ for $i = 1, \ldots, m-1$ and $\left(\frac{a_i}{a_m}\right)^2 \geq 1$ for $i = m+1, \ldots, n$. By direct calculation one can easily see that decreasing the absolute value of a component $i < m$ or increasing the absolute value of a component $i > m$ increases the Rayleigh quotient. Thus $\lambda(c) \leq \lambda(d)$. Finally, for nonzero c_i and c_j the increase of the weighted mean is nonzero if $a_i < a_j$.

For a decreasing sequence of a_i consider the increasing sequence $b_i := 1/a_i$ and the result from above to $c_i = b_i d_i$. □

The second lemma proves an inequality on various weighted means.

Lemma A.2. *For $c \in \mathbf{R}^k$ and $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_k$ let $\tau > \lambda_k$. Then we have*

$$\left(\sum_{i=1}^k c_i^2\right)\left(\sum_{i=1}^k \frac{c_i^2}{\lambda_i(\tau - \lambda_i)}\right)^2 \leq \left(\sum_{i=1}^k \frac{c_i^2}{\lambda_i}\right)^2 \left(\sum_{i=1}^k \frac{c_i^2}{(\tau - \lambda_i)^2}\right). \tag{A.1}$$

Proof. The proposition is equivalent to

$$\left(\frac{\sum_{i=1}^k \frac{c_i^2}{\lambda_i(\tau-\lambda_i)}}{\sum_{i=1}^k \frac{c_i^2}{\lambda_i}} \right)^2 \leq \frac{\sum_{i=1}^k \frac{c_i^2}{(\tau-\lambda_i)^2}}{\sum_{i=1}^k c_i^2}.$$

At first we show

$$\frac{\sum_{i=1}^k \frac{c_i^2}{\lambda_i(\tau-\lambda_i)}}{\sum_{i=1}^k \frac{c_i^2}{\lambda_i}} \leq \frac{\sum_{i=1}^k \frac{c_i^2}{\tau-\lambda_i}}{\sum_{i=1}^k c_i^2}, \tag{A.2}$$

or equivalently

$$\frac{\sum_{i=1}^k c_i^2}{\sum_{i=1}^k c_i^2/\lambda_i} \leq \frac{\sum_{i=1}^k \frac{c_i^2}{\tau-\lambda_i}}{\sum_{i=1}^k \frac{c_i^2}{\lambda_i(\tau-\lambda_i)}} = \frac{\sum_{i=1}^k \left(\frac{c_i}{\sqrt{\tau-\lambda_i}}\right)^2}{\sum_{i=1}^k \left(\frac{c_i}{\sqrt{\tau-\lambda_i}}\right)^2/\lambda_i}.$$

Both sides of this inequality are Rayleigh quotients of the form (2.6). The coefficients on the right–hand side are weighted by the monotone increasing sequence

$$1/\sqrt{\tau-\lambda_1}, \ldots, 1/\sqrt{\tau-\lambda_n},$$

so that application of Lemma A.1 proves (A.2). We conclude the proof by estimating the square of the right–hand side of (A.2) by applying the Cauchy–Schwarz inequality to the nominator

$$\left(\frac{\sum_{i=1}^k \frac{c_i^2}{\tau-\lambda_i}}{\sum_{i=1}^k c_i^2} \right)^2 = \frac{\left(\sum_{i=1}^k c_i \left(\frac{c_i}{\tau-\lambda_i}\right)\right)^2}{\left(\sum_{i=1}^k c_i^2\right)^2} \leq \frac{\sum_{i=1}^k \frac{c_i^2}{(\tau-\lambda_i)^2}}{\sum_{i=1}^k c_i^2}.$$

□

REFERENCES

[1] W.W. Bradbury and R. Fletcher. New iterative methods for solution of the eigenproblem. *Numer. Math.*, 9:259–267, 1966.

[2] J.H. Bramble. *Multigrid Method*. Pitman Research Notes in Mathematics Series. Longman Scientific & Technical, London, 1993.

[3] J.H. Bramble, J.E. Pasciak, and A.V. Knyazev. A subspace preconditioning algorithm for eigenvector/eigenvalue computation. *Adv. Comput. Math.*, 6:159–189, 1996.

[4] J.H. Bramble, J.E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comp.*, 55:1–22, 1990.

[5] F. Chatelin. *Eigenvalues of matrices*. John Wiley & Sons, Chichester, 1993.

[6] E.G. D'yakonov. Iteration methods in eigenvalue problems. *Math. Notes*, 34:945–953, 1983.

[7] E.G. D'yakonov and M.Y. Orekhov. Minimization of the computational labor in determining the first eigenvalues of differential operators. *Math. Notes*, 27:382–391, 1980.

[8] Y.T. Feng and D.R.J. Owen. Conjugate gradient methods for solving the smallest eigenpair of large symmetric eigenvalue problems. *Int. J. Numer. Methods Eng.*, 39:2209–2229, 1996.

[9] S.K. Godunov, V.V. Ogneva, and G.P. Prokopov. On the convergence of the modified method of steepest descent in the calculation of eigenvalues. *Amer. Math. Soc. Transl. Ser. 2*, 105:111–116, 1976.

[10] M.R. Hestenes and W. Karush. A method of gradients for the calculation of the characteristic roots and vectors of a real symmetric matrix. *J. Res. Nat. Bureau Standards*, 47:45–61, 1951.

[11] I. Ipsen. *A history of inverse iteration*, volume in Helmut Wielandt, Mathematische Werke, Mathematical Works, Vol. 2: Linear Algebra and Analysis, pages 464–472. Walter de Gruyter, Berlin, 1996.

[12] I. Ipsen. Computing an eigenvector with inverse iteration. *SIAM Rev.*, 39:254–291, 1997.

[13] A.V. Knyazev. Preconditioned eigensolvers -an oxymoron? . *Electron. Trans. Numer. Anal.*, 7:104–123, 1998.

[14] A.V. Knyazev and A.L. Skorokhodov. The rate of convergence of the method of steepest descent in a Euclidean norm . *USSR J. Comput. Math. and Math. Physics*, 28:195–196, 1988.

[15] A.V. Knyazev and A.L. Skorokhodov. On exact estimates of the convergence rate of the steepest ascent method in the symmetric eigenvalue problem . *Linear Algebra Appl.*, 154–156:245–257, 1991.
[16] D.E. Longsine and S.F. McCormick. Simultaneous Rayleigh–quotient minimization methods for $Ax = \lambda Bx$. *Linear Algebra Appl.*, 34:195–234, 1980.
[17] S.F. McCormick. A general approach to one–step iterative methods with application to eigenvalue problems. *J. Comp. Sys. Sci.*, 6:354–372, 1972.
[18] S.F. McCormick. Some convergence results on the method of gradients for $Ax = \lambda Bx$. *J. Comp. Sys. Sci.*, 13:213–222, 1976.
[19] S.F. McCormick and T. Noe. Simultaneous iteration for the matrix eigenvalue problem. *Linear Algebra Appl.*, 16:43–56, 1977.
[20] K. Neymeyr. A geometric theory for preconditioned inverse iteration applied to a subspace. Submitted to Math. Comput., 1999.
[21] K. Neymeyr. A posteriori error estimation for elliptic eigenproblems. Submitted to Math. Comput., 1999.
[22] K. Neymeyr. Why preconditioning gradient type eigensolvers? Submitted to SIAM J. Sci. Comp., 2000.
[23] B.N. Parlett. *The symmetric eigenvalue problem*. Prentice Hall, Englewood Cliffs New Jersey, 1980.
[24] W.V. Petryshyn. On the eigenvalue problem $Tu - \lambda Su = 0$ with unbounded and non–symmetric operators T and S. *Philos. Trans. Roy. Soc. Math. Phys. Sci.*, 262:413–458, 1968.
[25] G. Rodrigue. A gradient method for the matrix eigenvalue problem $Ax = \lambda Bx$. *Numer. Math.*, 22:1–16, 1973.
[26] B.A. Samokish. The steepest descent method for an eigenvalue problem with semi–bounded operators. *Izv. Vyssh. Uchebn. Zaved. Mat.*, 5:105–114, 1958.
[27] G.L.G. Sleijpen and H.A. van der Vorst. A Jacobi–Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 17:401–425, 1996.
[28] G.L.G. Sleijpen and F.W. Wubs. Effective preconditioning techniques for eigenvalue problems. Technical Report 1117, Universiteit Utrecht, Department of Mathematics, 1999.
[29] P.S. Vassilevski. Preconditioning nonsymmetric and indefinite finite element matrices. *J. Numer. Linear Algebra Appl.*, 1:59–76, 1992.
[30] J. Xu. A new class of iterative methods for nonselfadjoint or indefinite problems. *SIAM J. Numer. Anal.*, 29:303–319, 1992.
[31] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Rev.*, 34:581–613, 1992.
[32] X. Yang. A survey of various conjugate gradient algorithms for iterative solution of the largest/smallest eigenvalue and eigenvector of a symmetric matrix, Collection: Application of conjugate gradient method to electromagnetic and signal analysis. *Progress in electromagnetic research*, 5:567–588, 1991.
[33] H. Yserentant. On the multi–level splitting of finite element spaces for indefinite elliptic boundary value problems. *SIAM J. Numer. Anal.*, 23:581–595, 1986.
[34] H. Yserentant. Old and new convergence proofs for multigrid methods. In *Acta numerica*, pages 285–326. Cambridge Univ. Press, Cambridge, 1993.

MATHEMATISCHES INSTITUT DER UNIVERSITÄT TÜBINGEN, AUF DER MORGENSTELLE 10, 72076 TÜBINGEN, GERMANY.

E-mail address: neymeyr@na.uni-tuebingen.de