A geometric theory for preconditioned inverse iteration:
IV: On the fastest convergence cases

Klaus Neymeyr

Universität Rostock, Fachbereich Mathematik, Universitätsplatz 1,
18051 Rostock, Germany;

SUMMARY

In order to compute the smallest eigenvalue together with an eigenfunction of a self-adjoint elliptic partial differential operator one can use the preconditioned inverse iteration scheme, also called the preconditioned gradient iteration. For this iterative eigensolver estimates on the poorest convergence have been published by several authors. In this paper estimates on the fastest possible convergence are derived. To this end the convergence problem is reformulated as a two-level constrained optimization problem for the Rayleigh quotient. The new convergence estimates reveal a wide range between the fastest possible and the slowest convergence.

1. Introduction

Why derive estimates on the *fastest* possible convergence of an iterative eigensolver? This is a reasonable question in so far as the predicted convergence rate is determined by estimates on the *slowest* (or poorest) convergence.

However, convergence rate estimates for iterative eigensolvers for self-adjoint eigenvalue problems are sometimes unduly pessimistic! Prominent examples of solvers are iterations like the power method, the complementary inverse iteration or the Lanczos scheme. For all these iterations the convergence rate estimates depend on the eigenvalue distribution or, more specific, on quantities like the ratio of consecutive smallest/largest eigenvalues or on the spectral condition number of the matrix whose eigenvalues are to be computed. However, for certain iteration vectors these eigensolvers may converge much more rapidly than reflected by the (worst case) convergence estimates. There is a simple explanation for this quick convergence: the eigenvector expansion of the initial iterate might show only little contribution from eigenspaces which are responsible for the poorest convergence. In the extremal case of *no* contribution from certain eigenvectors, the iteration will take place in their orthogonal complement and any unfavorable influence of these eigenvalues disappears.

In this paper we analyze an *inexact* version of inverse iteration, called preconditioned inverse iteration or preconditioned gradient iteration. This eigensolver uses a preconditioner for convergence acceleration which is assumed to satisfy a certain quality constraint. Similarly to the existence of vectors associated with best/poorest convergence, there are also preconditioners which are associated with fastest or slowest convergence. Thus for a preconditioned eigensolver there are two factors which determine the convergence decisively: first of all the initial iteration vector and secondly the preconditioner.

Here we do not treat the important questions of how to find an appropriate initial vector and how to construct a favorable preconditioner in order to gain a fast converging iteration. Instead, our aim is to investigate the range between the *fastest* and the *slowest theoretically possible* convergence for a basic preconditioned eigensolver under reasonable assumptions on the preconditioner and on the initial vector. The practical question of how to accelerate the iteration due to an appropriate preconditioner is non-trivial; the present paper might prepare the ground for a better understanding of the whole problem and of the potential of preconditioned eigensolvers. Therefore the present analysis should be understood as a step towards an improved analytical understanding of practically successful preconditioned eigensolvers.

Sharp estimates on the slowest possible convergence have already been given in [9]. Hence, our present aim is to derive sharp estimates on the fastest convergence. These upper and lower estimates enclose a wide range between fastest and poorest convergence. The analysis shows that theoretically even one-step convergence to an eigenvector is possible. Such single-step convergence is an interesting phenomenon. It is totally different from that of iterative solvers like inverse iteration, which converges in infinitely many steps.

The paper is organized as follows: In Sec. 2 a basic preconditioned eigensolver is introduced and the problem to derive convergence estimates for this eigensolver is reformulated as a *two-level optimization problem.* In Sec. 3 the *inner* optimization problem to determine an optimal preconditioner is treated. The *outer* optimization problem on a level set of the Rayleigh quotient is analyzed in Sec. 4. Finally, all results are merged into the central convergence theorem in Sec. 5. Here we re-use arguments from [10, 11] which can partially be extended to local extrema; but we also point out certain non-trivial differences.

2. Preconditioned eigensolvers

Preconditioned eigensolvers are well suited for the partial solution of generalized eigenvalue problems which occur from a mesh discretization of a self-adjoint elliptic partial differential operator. Among other areas of application such eigenproblems appear in structural mechanics, see the references in [6, 8] for typical applications. Usually, only one or a modest number of the smallest eigenvalues together with the eigenvectors are to be determined. For instance these eigenpairs determine the lowest vibration modes of a mechanical system. The generalized matrix eigenproblem reads

$$Ax_i = \lambda_i M x_i \tag{1}$$

with $A \in \mathbb{R}^{n\times n}$ ($M \in \mathbb{R}^{n\times n}$) being called the discretization (mass) matrix and (x_i, λ_i) denoting an eigenpair. The matrices A and M are symmetric positive definite and, usually, very large and sparse. To simplify the representation we reduce (1) to the standard eigenproblem

$$Ax_i = \lambda_i x_i. \tag{2}$$

This reduction is justified by a change from the Euclidean inner product to the inner product induced by M; see [8]. In our setup there is no necessity to factor A or $A - \sigma I$, $\sigma \in \mathbb{R}$ and I the identity matrix. Such factorizations should be avoided because of storage and computation time limitations. Consequently, the application of an eigensolver which requires matrix factorizations (like the QR algorithm) appears to be impractical.

In this paper we consider the problem to determine only the smallest eigenvalue λ_1 together with an eigenvector x_1 of (2); see [1, 13] for a related subspace scheme. The "classical" derivation of *preconditioned eigensolvers* amounts to considering the eigenvalue problem as a minimization problem for the Rayleigh quotient

$$\lambda(x) = \frac{(x, Ax)}{(x, x)} \tag{3}$$

whose minimum is λ_1. Since the gradient $\nabla\lambda(x)$ is a multiple of the residual vector $Ax - \lambda(x)x$, a *gradient method* for minimizing the Rayleigh quotient maps a given iterate x to

$$x' = x - \omega(Ax - \lambda(x)x), \tag{4}$$

in order to attain $\lambda(x') < \lambda(x)$ for an appropriate choice of $\omega \in \mathbb{R}$. The aim is to construct a sequence of iterates converging to an eigenvector corresponding to the smallest eigenvalue. Unfortunately, it is well known that the convergence of the gradient method (4) depends on the mesh size and therefore on the number of unknowns [3]. Thus the gradient scheme cannot be considered as an effective solver for mesh eigenproblems.

Preconditioning can assure *grid-independent* convergence. A preconditioner $B^{-1} \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix which approximates the inverse of A. Especially for A being a mesh discretization of an elliptic partial differential operator, the preconditioner can be characterized by a spectral equivalence

$$\gamma_0(x, Bx) \leq (x, Ax) \leq \gamma_1(x, Bx) \tag{5}$$

for real positive constants γ_0 and γ_1. We assume an optimally scaled preconditioner (such a scaling can often be guaranteed implicitly, cf. [8]), i.e., we have instead of (5)

$$\|I - B^{-1}A\|_A \leq \gamma, \qquad 0 \leq \gamma < 1 \tag{6}$$

with γ controlling the quality of B^{-1}. Here, we do not raise the issue of how to construct such preconditioners satisfying (6), but refer to the references in [1, 3, 6].

A *basic* preconditioned eigensolver can be constructed from (4) just by premultiplying the residual vector by B^{-1}. This has been interpreted as a change of the underlying geometry in a way which accelerates convergence [3, 14]. Thus the new iterate $x' \in \mathbb{R}^n$ is given by

$$x' = x - B^{-1}(Ax - \lambda(x)x). \tag{7}$$

There is a vast literature on the convergence theory of (7), see the references in [3, 8]. While the older analysis has resulted in non-sharp or, at best, in asymptotically sharp convergence estimates, one can derive sharp convergence estimates using an alternative derivation of (7), see [10, 11]. The key idea is to interpret (7) as an approximate variant of *inverse iteration.* Inverse iteration for A amounts to solving the linear system

$$A\bar{x} = \lambda(x)x \tag{8}$$

for the new iterate $\bar{x}$; in contrast to the standard representation of inverse iteration the right-hand side is additionally scaled with $\lambda(x)$. Approximate solution of (8) using preconditioning leads to the *error propagation equation*

$$x' - \lambda(x)A^{-1}x = (I - B^{-1}A)(x - \lambda(x)A^{-1}x) \tag{9}$$

with x' approximating the exact solution $\bar{x} = \lambda(x)A^{-1}x$. Eq. (9) is not only a reformulation of (7), but establishes a relation between *preconditioned gradient eigensolvers* and approximate inverse iteration or *preconditioned inverse iteration.* A favorable property of (9) is the appearance of the error propagation matrix $I - B^{-1}A$, which allows a new approach to the analysis. A convergence analysis exploiting the structure of (9) is contained in [10, 11]. In this paper the very same techniques are used to derive estimates on the fastest possible convergence.

2.1. Convergence analysis as an optimization problem

Our aim is to compute the smallest eigenvalue λ_1 of (2) together with an eigenvector by using (7). As introduced above, this partial eigenvalue problem is considered as a minimization problem for the Rayleigh quotient. Thus the task to derive estimates on the fastest possible convergence of (7) can be reformulated as a *two-level optimization problem.* The two levels are as follows:

1. *Inner optimization problem:* For given $\gamma \in [0,1)$ let

$$\mathcal{B}_\gamma := \{B^{-1} \in \mathbb{R}^{n\times n};\ B \text{ symmetric positive definite, } \|I - B^{-1}A\|_A \leq \gamma\}$$

be the set of admissible preconditioners. The optimization problem consists in finding the specific $B^{-1} \in \mathcal{B}_\gamma$ which minimizes $\lambda(x')$ with x' by (7). This problem is analyzed in Sec. 3.

2. *Outer optimization problem:* Consider the level set

$$\mathcal{L}(\lambda) := \{x \in \mathbb{R}^n;\ \lambda(x) = \lambda\}$$

of vectors whose Rayleigh quotient equals a real number λ between the smallest and the largest eigenvalue of A. Minimization is to be done with respect to the level set $\mathcal{L}(\lambda)$, i.e., to find that $x \in \mathcal{L}(\lambda)$ which minimizes the Rayleigh quotient $\lambda(x')$ of the new iterate. The analysis is presented in Sec. 4.

The optimal choices from both $\mathcal{B}_\gamma$ and from $\mathcal{L}(\lambda)$ lead to the fastest possible convergence and result in the smallest attainable Rayleigh quotient

$$\min_{x\in\mathcal{L}(\lambda)} \quad \min_{B^{-1}\in\mathcal{B}_\gamma} \lambda(x - B^{-1}(Ax - \lambda x)). \tag{10}$$

Note that *exact* preconditioning solves (8) exactly, i.e., $B = A$ results in $x' = \bar{x}$. In contrast to this, the *optimal* preconditioner minimizes (10). This makes a fundamental difference between optimal preconditioning for linear systems and for eigenvalue problems.

2.2. Geometric representation and change of the basis

Lemma 2.1 provides a geometric description of the constraint $B^{-1} \in \mathcal{B}_\gamma$ and yields a more convenient reformulation of the optimization problem (10).

Lemma 2.1. *Let $x \in \mathbb{R}^n$, $x \neq 0$, and let*

$$B_\gamma(x) := \{\lambda A^{-1}x + y;\ \ y \in \mathbb{R}^n,\ \|y\|_A \leq \gamma\|(I - \lambda A^{-1})x\|_A\},$$

which is a ball with respect to the norm induced by A with the center $\bar{x} = \lambda A^{-1}x$, i.e., the solution of (8). Then the mapping E_x given by

$$E_x : \mathcal{B}_\gamma \to B_\gamma(x) :\ \ B^{-1} \mapsto x' = x - B^{-1}(Ax - \lambda x)$$

is a surjection.

The proof follows from Lemmata 2.2 and 2.3 of [10]. Therefore the inner optimization problem of Sec. 2.1 is equivalent to finding the minimum of $\lambda(\cdot)$ on the ball $B_\gamma(x)$. We transform this problem in a more convenient form by introducing a basis of A-orthonormal eigenvectors of A, see Sec. 2 in [10]. The initial basis ("x-basis") is mapped to a new basis (briefly "c-basis") by

$$c = \Lambda^{1/2}X^T x. \tag{11}$$

Therein, the orthogonal matrix X diagonalizes A, i.e., $X^TAX = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with $X^TX = I$. The eigenvalues λ_i are assumed to be simple (see Sec. 3 in [10] for a treatment of

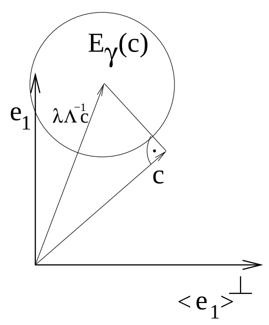

Figure 1. *One-step convergence.*

the multiple eigenvalue case) and the corresponding eigenvector e_i is the ith column of the identity matrix I. Then the c-basis representation of the Rayleigh quotient of a vector $d \in \mathbb{R}^n$ reads

$$\lambda(d) = \frac{(d, d)}{(d, \Lambda^{-1}d)}. \tag{12}$$

Reformulation of the two-level optimization problem (10) results in

$$\min_{c \in L(\lambda)} \quad \min_{d \in E_\gamma(c)} \quad \lambda(d). \tag{13}$$

Therein the c-basis representation of the level set using (12) is

$$L(\lambda) := \{c \in \mathbb{R}^n;\ \lambda(c) = \lambda\}. \tag{14}$$

Moreover, the ball

$$E_\gamma(c) := \{\lambda\Lambda^{-1}c + z;\ \ z \in \mathbb{R}^n,\ \|z\| \le \gamma\|(I - \lambda\Lambda^{-1})c\|\ \} \tag{15}$$

is the c-basis representation of $B_\gamma(x)$; $\|\cdot\|$ denotes the Euclidean norm.

Next we make certain non-restrictive assumptions on $c \in \mathbb{R}^n$; see Sec. 4 in [10] for a justification.

Assumption 2.2. *For given $\lambda \in (\lambda_1, \lambda_n)$ the vector $c \in \mathbb{R}^n$ satisfies*

1. *$c \in L(\lambda)$ and $\|c\| = 1$,*
2. *c is not equal to any of the unit vectors e_i, $i = 1, \ldots, n$,*
3. *$c \ge 0$ componentwise.*

3. The inner optimization problem: Optimal preconditioning

In this section the inner optimization problem of (13), i.e.,

$$\min_{d \in E_\gamma(c)} \quad \lambda(d) \tag{16}$$

is to be solved. As the Rayleigh quotient is invariant with respect to a scaling of its argument, we can alternatively consider the minimization with respect to the set $C_\gamma(c)$ being the smallest circular cone enclosing $E_\gamma(c)$ and with vertex at the origin

$$C_\gamma(c) := \{\xi d;\ d \in E_\gamma(c),\ \xi > 0\}. \tag{17}$$

3.1. Localization of minima

The preconditioned eigensolver (7) exhibits the surprising property that for certain $c \in L(\lambda)$ even *one-step convergence* may occur; i.e. in only one iteration the eigenvector e_1 corresponding to the smallest eigenvalue λ_1 can be reached. A corresponding geometric setup in $\mathbb{R}^2$ is shown in Figure 1. One-step convergence is possible if the cone $C_\gamma(c)$ contains the eigenvector e_1. Lemma 3.1 provides a condition for one-step convergence.

Lemma 3.1. *Let $c \in \mathbb{R}^n$ be given according to Assumption 2.2. Then one-step convergence, i.e., $e_1 \in C_\gamma(c)$, may occur if and only if*

$$c_1 \geq \frac{\lambda_1}{\lambda}\Big(\|\lambda\Lambda^{-1}c\|^2 - \gamma^2\|(I-\lambda\Lambda^{-1})c\|^2\Big)^{1/2}. \tag{18}$$

Proof. The acute angle χ between e_1 and $\lambda\Lambda^{-1}c$ is given by

$$\cos\chi = \frac{\lambda\lambda_1^{-1}c_1}{\|\lambda\Lambda^{-1}c\|}.$$

For the opening angle φ of $C_\gamma(c)$ by using the orthogonal decomposition from Thm. 4.3 in [10] one obtains that

$$\cos^2\varphi = \frac{\|\lambda\Lambda^{-1}c\|^2 - \gamma^2\|(I-\lambda\Lambda^{-1})c\|^2}{\|\lambda\Lambda^{-1}c\|^2}.$$

Then $\chi \leq \varphi$ yields $\lambda\lambda_1^{-1}c_1 \geq \big(\|\lambda\Lambda^{-1}c\|^2 - \gamma^2\|(I-\lambda\Lambda^{-1})c\|^2\big)^{1/2}$ which proves (18). □

Inequality (18) is not a hard condition and is further weakened for increasing γ. Increasing of γ results in a larger set $\mathcal{B}_\gamma$ and, due to Lemma 2.1, in a larger ball B_γ. The limit cone $\lim_{\gamma\to 1} C_\gamma(c)$ contains e_1, if and only if

$$c_1 \geq \frac{\lambda_1}{\lambda}, \tag{19}$$

which follows from (18) together with $\|\lambda\Lambda^{-1}c\|^2 = \|c\|^2 + \|(I-\lambda\Lambda^{-1})c\|^2$. See Fig. 3 in Sec. 3.2 for an example satisfying condition (19).

From now on we restrict our attention to the non-trivial case $e_1 \notin C_\gamma(c)$, i.e.,

$$\min_{d\in C_\gamma(c)} \lambda(d) > \lambda_1.$$

Our next aim is to locate points of extrema of the Rayleigh quotient on $E_\gamma(c)$ (or equivalently $C_\gamma(c)$) by analyzing its local behavior. The following Lemma 3.2 shows that the minima are taken on the $(n-2)$-dimensional manifold

$$\mathcal{M} = (\partial C_\gamma(c)) \cap E_\gamma(c), \tag{20}$$

with $\partial C_\gamma(c)$ denoting the boundary of $C_\gamma(c)$. The manifold $\mathcal{M}$ is characterized by the constraints (22a) and (22b).

Lemma 3.2. *Let c satisfy Assumption 2.2 and $e_1 \notin C_\gamma(c)$. Then*

$$\arg\min \lambda(E_\gamma(c)) \ \subset \ \mathcal{M}, \tag{21}$$

with arg min *denoting the set of minimum points. For any $w \in \arg\min\lambda(E_\gamma(c))$ it holds that*

$$(w, w-\lambda\Lambda^{-1}c) = 0, \tag{22a}$$

$$\|\lambda\Lambda^{-1}c\|^2 = \|w\|^2 + \|w-\lambda\Lambda^{-1}c\|^2, \tag{22b}$$

$$\|w-\lambda\Lambda^{-1}c\| = \gamma\|(I-\lambda\Lambda^{-1})c\|, \tag{22c}$$

Proof. The minimum (16) cannot be taken in the interior of the cone $C_\gamma(c)$ as $e_1 \notin C_\gamma(c)$, and all other stationary points of $\lambda(\cdot)$ on $C_\gamma(c)$ are saddle points, see Lemmata 4.1 and 4.2 of [10]. Hence, (21) holds. The orthogonality (22a) and the decomposition (22b) is true for any $w \in \mathcal{M}$, since the tangent plane to $\partial C_\gamma(c)$ in w is also a tangent plane to $\partial E_\gamma(c)$ in w. Finally, by (22c) the radius $\|w - \lambda\Lambda^{-1}c\|$ of $E_\gamma(c)$ is expressed as γ times the radius of the maximal ball $E_1(c)$. □

Let us now determine those w from the manifold $\mathcal{M}$ in which the Rayleigh quotient takes a relative extremum. We apply the method of Lagrange multipliers in order to derive a necessary condition on a local extremum of $\lambda(\cdot)\mid_{\mathcal{M}}$. For given c the norm $\|w\|$ is a constant on $\mathcal{M}$, i.e. by (22b) and (22c) it holds that

$$(w, w) = \|\lambda\Lambda^{-1}c\|^2 - \gamma^2\|(I - \lambda\Lambda^{-1})c\|^2, \qquad \forall w \in \mathcal{M}.$$

Hence extrema of $\lambda(w)$ and those of the quadratic function $(w, \Lambda^{-1}w)$, $w \in \mathcal{M}$, are taken in the same arguments. Thus the Lagrange function with respect to the constraints (22a) and (22b) determining $\mathcal{M}$ with the Lagrange multipliers μ and ν reads

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

We obtain from $\nabla_w\mathcal{L} = 0$ a condition on w

$$2(\Lambda^{-1} + (\mu + \nu)I)w = \nu\lambda\Lambda^{-1}c. \tag{23}$$

An equivalent condition can be derived by noticing that the gradient $\nabla\lambda(w)$ in a local extremum on $E_\gamma(c)$ is orthogonal to the tangent plane to $E_\gamma(c)$ in w [5].

Note that $\nu \neq 0$ in (23). Otherwise any solution w of (23) would be a multiple of a unit vector e_i. Here we do not present the somewhat technical proof that in unit vectors e_i, $i \geq 2$, the Rayleigh quotient never takes a minimum on $E_\gamma(c)$. We refer to Lemma A.1 in [11] whose arguments can be extended to minima.

In order to solve (23) for w, the diagonal matrix $D := \Lambda^{-1} + (\mu + \nu)I$ has to be inverted. If in w a local maximum is taken, then D is invertible as shown by Thm. 4.8 in [10]. But this is *not* always the case for minimum points; see Sec. 3.2 for a numerical example. Problems occur if $c_1 = 0$. Nevertheless, Lemma 3.3 guarantees $D_{ii} \neq 0$ for $i > 1$.

Lemma 3.3. *On the assumptions of Lemma 3.2 let $w \in \arg\min \lambda(E_\gamma(c))$. If $c_k > 0$, then*

$$w_k = \frac{\lambda\nu}{2 + 2\lambda_k(\mu + \nu)}\, c_k > 0. \tag{24}$$

for $k = 1, \ldots, n$. If $c_k = 0$, then $w_k = 0$ for $k = 2, \ldots, n$. Finally, $c_1 = 0$ does not imply $w_1 = 0$, see Sec. 3.2.

The proof of Lemma 3.3 follows along the lines of Lemma 4.7 in [10]: For non-zero c_k the representation (24) immediately follows from (23). Then it is shown that $w_k \neq 0$ together with $c_k = 0$ can hold for not more than a single k. By slightly adapting the arguments of Lemma 4.7 in [10], one can show that only $w_1 \neq 0$ together with $c_1 = 0$ can occur; see also [12].

In the following we assume $c_1 \neq 0$ which is the case if $\lambda(c) < \lambda_2$. The latter assumption is often used, e.g., in the classical convergence analysis of preconditioned gradient methods [2, 4]. Thm. 3.4 shows that for each $\gamma \in [0, 1)$ the minimum of $\lambda(E_\gamma(c))$ is taken in a unique point. Moreover, the set of all minimum points for all $\gamma \in [0, 1)$ is a curve parametrized in α.

Theorem 3.4. *On the assumptions of Lemma 3.2 and if $c_1 > 0$, then the minimum $\lambda(E_\gamma(c))$ is taken in*

$$w[\alpha] = \beta(\alpha I + \Lambda)^{-1}c, \tag{25}$$

for a unique real number $\alpha \in (-\lambda_1, 0]$. Therein $\beta = \beta[\alpha]$ is given by

$$\beta[\alpha] = \frac{(\lambda\Lambda^{-1}c, (\alpha I + \Lambda)^{-1}c)}{((\alpha I + \Lambda)^{-1}c, (\alpha I + \Lambda)^{-1}c)} > 0.$$

Then all Rayleigh quotients $\lambda(w[\alpha])$ for $\gamma \in [0, 1)$ form a subinterval of the image of the strictly monotone increasing function

$$\rho : (-\lambda_1, 0] \to (\lambda_1, \lambda(\Lambda^{-1}c)] : \alpha \mapsto \lambda(w) = \lambda((\alpha I + \Lambda)^{-1}c). \tag{26}$$

Proof. From (23) and Lemma 3.3 any $w \in \arg\min \lambda(E_\gamma(c))$ can be written in the form (25) for certain $\alpha, \beta \in \mathbb{R}$. The coefficients α and β depend on $\gamma \in [0, 1)$.

First it is shown that $\beta > 0$ and $\alpha > -\lambda_1$. For $w = \beta(\alpha I + \Lambda)^{-1}c$ it holds $\beta/(\alpha + \lambda_i) > 0$ for any nonzero component c_i by Lemma 3.3. If $\beta < 0$, then $\alpha < -\lambda_l$ (with l being 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, $\lambda(w) > \lambda(c)$, which contradicts the monotone decrease of the Rayleigh quotient or convergence of (7), see [11]. Thus $\beta > 0$ and $\alpha + \lambda_1 > 0$. The explicit form of $\beta > 0$ can be gained from (22a).

In order to show that ρ is a strictly monotone increasing function, note that for $\alpha > -\lambda_1$ the diagonal matrix $(\alpha I + \Lambda)$ is invertible. Let $-\lambda_1 < \alpha_1 < \alpha_2$ be given and define $w^{(1)} := (\alpha_1 I + \Lambda)^{-1}c$ and $w^{(2)} := (\alpha_2 I + \Lambda)^{-1}c$. Then for $i = 1, \ldots, n$

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

The positive coefficients $(\alpha_2 + \lambda_1)/(\alpha_1 + \lambda_1), \ldots, (\alpha_2 + \lambda_n)/(\alpha_1 + \lambda_n)$ form a strictly monotone decreasing sequence. Thus Lemma A.1 in [10] shows that ρ is a strictly monotone increasing function. Furthermore, it holds

$$\lim_{\alpha \to -\lambda_1} \lambda((\alpha I + \Lambda)^{-1}c) = \lambda_1.$$

Uniqueness of α and of the minimum point $w[\alpha]$ follows from the monotonicity of (26) and the fact that $\lambda((\alpha I + \Lambda)^{-1}c) > \lambda(\Lambda^{-1}c)$ for $\alpha > 0$, which contradicts $(\alpha I + \Lambda)^{-1}c$ being a minimum point. □

Eq. (25) provides a single parameter representation of the minimum points for $\gamma \in [0, 1)$. A challenging problem is to derive a *re-parametrization* of $w[\alpha]$ as a function of $\gamma \in [0, 1)$. Such a representation $w[\gamma]$ would allow a considerable simplification of our convergence analysis. Unfortunately, the problem to determine α as a function of γ is not easy to tackle. In the $\mathbb{R}^n$ this requires the solution of a polynomial of degree $2n-2$ in α. A solution for $n = 2$ is given in Sec. 3.1.4 of [12].

3.2. Bifurcation of the minimum curve

If $c_1 = 0$, then $\rho(\alpha)$ by (26) cannot represent all minimum points, since then

$$\min_{\alpha \in (-\lambda_1, 0]} \rho(\alpha) \geq \lambda_2,$$

but it may hold $\min \lambda(E_\gamma(c)) < \lambda_2$. A discussion of the limit $c_1 \to 0$ can provide insight into the case $c_1 = 0$. One finds that as long as

$$\lambda((\Lambda - \lambda_1 I)^{+}c) \leq \min \lambda(E_\gamma(c)),$$

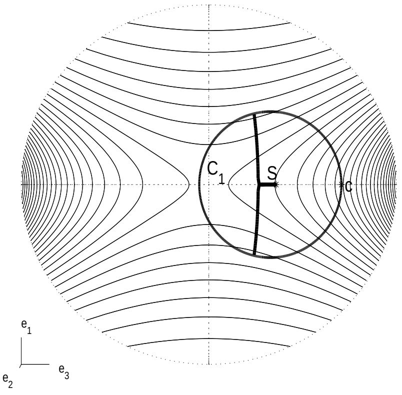

Figure 2. *Bifurcation of curve $S(\alpha)$ of minimum points on $E_\gamma(c)$, $\gamma \in [0,1]$.*

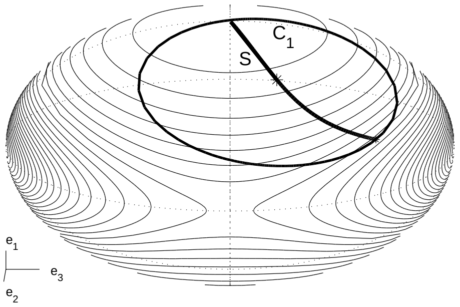

Figure 3. *Curve S of extremum points on $E_\gamma(c)$,$\gamma \in [0,1]$.*

where + denotes the pseudo-inverse, the form of the minimum points is determined by Thm. 3.4. Beyond the bound $\lambda((\Lambda - \lambda_1 I)^+ c)$ the minimum points have the form (aside from scaling)

$$\pm\vartheta e_1 + (\Lambda - \lambda_1 I)^+ c \tag{27}$$

for suitable $\vartheta \geq 0$.

A numerical example in $\mathbb{R}^3$ (the smallest nontrivial dimension) is given in Fig. 2. We take $\Lambda = \mathrm{diag}(2, 5, 13)$. The unit sphere is projected along the e_2 axis, and isocurves of the Rayleigh quotient are drawn for $\lambda = \lambda_1 + (\lambda_3 - \lambda_1)\frac{i}{30}$ with $i = 1, \ldots, 29$. For $c = (0, 1/\sqrt{2}, 1/\sqrt{2})^T$ the intersection of $C_1(c)$ with the unit sphere is shown as the bold circle C_1. The curve S of minimum points for $\gamma \in [0,1]$ (bold T-shaped curve) starts at the center of the circle ($\gamma = 0$) and bifurcates at $\gamma \approx 0.248$ in $w = (\Lambda - \lambda_1 I)^+ c$. The branches are of the form (27).

Fig. 3 illustrates the curve S of minimum and maximum points for $\gamma \in [0,1]$. Once again $\Lambda = \mathrm{diag}(2, 5, 13)$ but $c = (3, 5, 5)^T/\sqrt{59}$. Now, $\alpha \in (-\lambda_1, \infty)$ and the smooth curve

$$S(\alpha) = \frac{(\alpha I + \Lambda)^{-1} c}{\|(\alpha I + \Lambda)^{-1} c\|} \tag{28}$$

starts at the north pole ($\alpha \to -\lambda_1$), runs through the axis $\lambda\Lambda^{-1}c$ of the cone for $\alpha = 0$ and finally reaches in the initial vector c for $\alpha \to \infty$. Therein, all $\alpha < 0$ correspond to minimum points whereas $\alpha > 0$ gives the representation of maximum points. For this example the condition (19) is fulfilled since $0.391 \approx c_1 > \lambda_1/\lambda \approx 0.387$. Hence, the eigenvector e_1 is contained in $C_1(c)$.

4. The outer optimization problem on $L(\lambda)$

In this section the outer minimization problem of (13)

$$\min_{c\in L(\lambda)} \lambda(w[c])$$

with $w[c] := \arg\min_{d\in E_\gamma(c)} \lambda(d)$ is treated. In Sec. 4.1 we derive certain extremal properties of $C_\gamma(c)$ on $L(\lambda)$. In Sec. 4.2 convergence estimates in $\mathbb{R}^2$ are presented which form the basis for the main convergence theorem in Sec. 5.

4.1. Extrema of $C_\gamma(c)$ on $L(\lambda)$

Lemma 4.1 is a generalization of Thm. 2.1 in [11]. Extrema of $\|\nabla\lambda(c)\|$ are shown to be taken in two-dimensional invariant subspaces.

Lemma 4.1. *Let $\lambda = \lambda(c) \in (\lambda_1, \lambda_n)$. Then for the Euclidean norm of the gradient*

$$\nabla\lambda(c) = \frac{2}{(c, \Lambda^{-1}c)}(I - \lambda\Lambda^{-1})c$$

it holds:

1. *If $\lambda = \lambda_i$, then $\|\nabla\lambda(e_i)\| = 0$ is an absolute minimum. If $\lambda_i < \lambda < \lambda_{i+1}$, then the minimum of $\|\nabla\lambda(c)\|$ on $L(\lambda)$ is taken in a vector of the form*

$$c_{i,i+1} := (0, \dots, 0, c_i, c_{i+1}, 0, \dots, 0)^T \in L(\lambda), \tag{29}$$

   having exactly the two non-zero components c_i and c_{i+1}.
2. *The maximum of $\|\nabla\lambda(c)\|$ on $L(\lambda)$ is taken in a vector of the form*

$$c_{1,n} = (c_1, 0, \dots, 0, c_n)^T \in L(\lambda). \tag{30}$$

If c satisfies Assumption 2.2, then the components of (29) and (30) are uniquely determined, see (41).

Proof. The method of Lagrange multipliers for

$$\mathcal{L}(c, \mu, \nu) = \|(I - \lambda\Lambda^{-1})c\|^2 + \mu(\|c\|^2 - 1) + \nu((c, \Lambda^{-1}c) - \lambda^{-1}) \tag{31}$$

yields a necessary condition for a constrained local extremum of $\|\nabla\lambda(c)\|$ on $L(\lambda)$; see Thm. 2.1 in [11] for the details. One finally obtains the Temple-type inequality

$$4\lambda^2(\frac{\lambda}{\lambda_i} - 1)(1 - \frac{\lambda}{\lambda_{i+1}}) \ \le\ \|\nabla\lambda(c)\|^2 \ \le\ 4\lambda^2(\frac{\lambda}{\lambda_1} - 1)(1 - \frac{\lambda}{\lambda_n}). \tag{32}$$

The lower bound is taken in $c_{i,i+1}$ and the upper bound in $c_{1,n}$. □

We note that the left inequality in (32) has already been given, e.g., in Chap.9, §3 of [3]; the right inequality can be derived similarly.

These extrema of $\|\nabla\lambda(c)\|$ are closely related with extremal properties of the geometry of $C_\gamma(c)$. We introduce the *opening angle* $\varphi_\gamma(c)$ of the circular cone $C_\gamma(c)$ by

$$\varphi_\gamma(c) := \sup_{z\in C_\gamma(c)} \arccos(\frac{\lambda\Lambda^{-1}c}{\|\lambda\Lambda^{-1}c\|}, \frac{z}{\|z\|}). \tag{33}$$

The complementary *shrinking angle* $\varphi_\gamma(c)$ can be defined as

$$\psi_\gamma(c) := \varphi_1(c) - \varphi_\gamma(c).$$

The shrinking angle turns out to be relevant as the action of (7) can be understood as a shrinking of the initial cone $C_1(c)$ in the following sense. The iterate c is the maximum point of the Rayleigh quotient on the surface of $C_1(c)$, whereas for $\gamma < 1$ the global extrema are taken on the surface of the shrinked cone $C_\gamma(c)$ (aside from $e_1 \in C_\gamma(c)$). Lemma 4.2 reveals a close relation between $\|\nabla\lambda(c)\|$ and $\varphi_\gamma(c)$, $\psi_\gamma(c)$; cf. Lemmata 2.2 and 2.3 in [11].

Lemma 4.2. *Let $\lambda \in (\lambda_1, \lambda_n)$ and $\gamma \in [0, 1]$.*

1. *The trivial minimum $\varphi_\gamma(c) = 0$ ($\psi_\gamma(c) = 0$) can only be taken if $\gamma = 0$ ($\gamma = 1$) or if $\lambda = \lambda_i$ and $c = e_i$ for $i = 2, \ldots, n-1$. If $\lambda_i < \lambda < \lambda_{i+1}$, then the angles $\varphi_\gamma(c)$ and $\psi_\gamma(c)$ take their minima on $L(\lambda)$ in $c_{i,i+1}$.*
2. *The angles $\varphi_\gamma(c)$ and $\psi_\gamma(c)$ take their maxima on $L(\lambda)$ in $c_{1,n}$.*

The proof of Lemma 4.2 immediately follows from extending the proofs of Lemmata 2.2 and 2.3 in [11] to maxima.

Lemma 4.2 allows to analyze the dependence of the Rayleigh quotient on the opening angle φ_γ within the plane

$$P_{c,w} := \text{span}\{\lambda\Lambda^{-1}c, w\}, \tag{34}$$

through the minimum point w by (25) and $\lambda\Lambda^{-1}c$.

Now parametrize the unit circle in $P_{c,w}$ by $z(\varphi)$ so that $\varphi = \measuredangle(z(\varphi), \lambda\Lambda^{-1}c)$ and $z(\varphi^*) = w/\|w\|$ with $\varphi^* < \pi$. To express the angle dependence of the Rayleigh quotient in $P_{c,w}$ we define

$$\lambda_{c,w}(\varphi) := \lambda(z(\varphi)).$$

If c satisfies Assumption 2.2 and $c_1 > 0$, then for the derivative of the Rayleigh quotient w.r.t. to φ in w it holds that

$$|\frac{d\lambda_{c,w}}{d\varphi}(\varphi^*)| = \|\nabla\lambda(\frac{w}{\|w\|})\|, \tag{35}$$

whose proofs can literally be taken from Lemma 2.5 in [11].

Now define $\underline{\lambda}(c, \varphi)$ as the minimum of the Rayleigh quotient on $C_\gamma(c)$ having the opening angle $\varphi = \varphi_\gamma$, i.e.,

$$\underline{\lambda}(c, \varphi) := \inf \lambda(C_{\gamma(\varphi)}(c)),$$

for $\varphi \in [0, \arccos((c, \Lambda^{-1}c)/(\|c\|\|\Lambda^{-1}c\|)]$.

Lemma 4.3 discloses the identity of the derivatives $(d\underline{\lambda}(c, \varphi)/d\varphi)$ and $(d\lambda_{c,w}(\varphi)/d\varphi)$ within minimum points.

Lemma 4.3. *On the assumptions of Thm. 3.4 let w be a minimum point which encloses the angle $\varphi^* = \measuredangle(\lambda\Lambda^{-1}c, w)$ with the axis $\lambda\Lambda^{-1}c$ of $C_\gamma(c)$. Then it holds*

$$|\frac{d\underline{\lambda}}{d\varphi}(c, \varphi^*)| = |\frac{d\lambda_{c,w}}{d\varphi}(\varphi^*)| = \|\nabla\lambda(\frac{w}{\|w\|})\|. \tag{36}$$

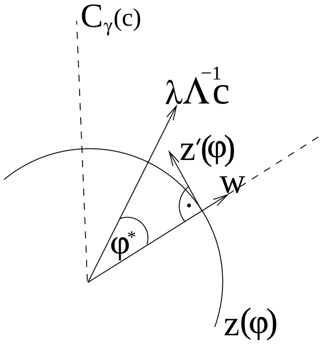

Figure 4. *2D cross-section $P_{c,w}$.*

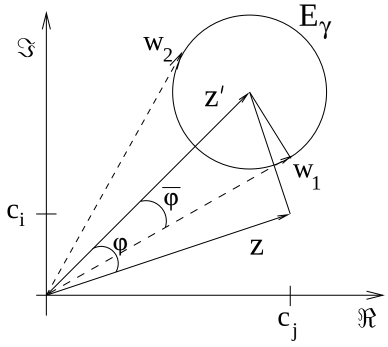

Figure 5. *Geometry in $\mathbb{C}$.*

Proof. Both $\lambda_{c,w}(\varphi)$ and $\underline{\lambda}(c,\varphi)$ are continuously differentiable in φ. By definition, $\lambda_{c,w}(\varphi)$ dominates $\underline{\lambda}(c,\varphi)$ for $\varphi \in [0,\varphi_1]$ so that

$$\underline{\lambda}(c,\varphi) \leq \lambda_{c,w}(\varphi) \quad \text{ and } \quad \lambda_{c,w}(\varphi^*) = \underline{\lambda}(c,\varphi^*),$$

where the last identity results from the fact that both functions coincide in φ^* belonging to the minimum point $w/\|w\|$. Since $\lambda_{c,w}(\varphi) - \underline{\lambda}(c,\varphi)$ is a positive differentiable function taking its minimum in φ^*, we conclude

$$\frac{d\lambda_{c,w}}{d\varphi}(\varphi^*) = \frac{d\underline{\lambda}}{d\varphi}(c,\varphi^*).$$

The proposition follows with (35). □

4.2. Mini-dimensional analysis in $\mathbb{C}$

In Sec. 4.1 it has been shown that several quantities which define the geometry of (7) take their extremal values in 2D invariant subspaces. Hence, not surprisingly, extremal convergence emerges in these 2D subspaces. In preparation of the main convergence theorem in Sec. 5, Thm. 4.4 gives convergence estimates in 2D on the fastest and on the slowest convergence. This *mini-dimensional* analysis is fairly different from that in [10]. It yields in the complex plane a more structured representation of the convergence estimates.

Theorem 4.4. *Let $\Lambda = \operatorname{diag}(\lambda_i, \lambda_j)$, $\lambda_i < \lambda_j$ and $c = (c_i, c_j)^T \in \mathbb{R}^2$ with $\lambda = \lambda(c)$.*
Then the maximal Rayleigh quotient on $E_\gamma(c)$ reads

$$\lambda(w_1) = \lambda^+(\lambda_i, \lambda_j, \lambda, \gamma) \tag{37}$$

with $w_1 \in \arg\max \lambda(E_\gamma(c))$.
In the trivial case $e_1 \in C_\gamma(c)$ the minimum of $\lambda(E_\gamma(c))$ equals λ_1. Otherwise,

$$\lambda(w_2) = \lambda^-(\lambda_i, \lambda_j, \lambda, \gamma) \tag{38}$$

is the minimum of the Rayleigh quotient with $w_2 \in \arg\min \lambda(E_\gamma(c))$. The functions $\lambda^\pm$ are given by

$$\lambda^\pm(\lambda_i, \lambda_j, \lambda, \gamma) = \left\{ \frac{1}{\lambda_i}\left(c_i\sqrt{1-(\rho^\pm)^2} + c_j\rho^\pm\right)^2 + \frac{1}{\lambda_j}\left(c_j\sqrt{1-(\rho^\pm)^2} - c_i\rho^\pm\right)^2 \right\}^{-1} \tag{39}$$

with

$$\rho^{\pm} = \xi \left(\sqrt{1-\gamma^2\xi^2} \mp \gamma\sqrt{1-\xi^2} \right) \tag{40}$$

and

$$\xi = \sqrt{\frac{(\lambda-\lambda_i)(\lambda_j-\lambda)}{\lambda(\lambda_i+\lambda_j-\lambda)}}, \qquad c_i = \sqrt{\frac{\lambda_i(\lambda_j-\lambda)}{\lambda(\lambda_j-\lambda_i)}}, \qquad c_j = \sqrt{\frac{\lambda_j(\lambda-\lambda_i)}{\lambda(\lambda_j-\lambda_i)}}. \tag{41}$$

Proof. According to Lemma 3.2 (case of minima) and Thm. 4.3 in [10] (case of maxima) it is clear that $(\partial C_\gamma(c)) \cap E_\gamma(c) \subset \mathbb{R}^2$ contains only two elements, i.e. the maximum point and the minimum point of the Rayleigh quotient on $E_\gamma(c)$. Our analysis to determine these extrema is based on an alternative approach compared to the construction used in Thm. 5.1 in [10]. Here the plane $\mathbb{R}^2$ is mapped to the complex plane according to

$$\tau : \mathbb{R}^2 \to \mathbb{C} : \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \mapsto y_2 + iy_1.$$

The Rayleigh quotient (12) is a scaling-invariant function. Thus $\lambda(\tau^{-1}(\theta\tau(y))) = \lambda(y)$ for all $\theta \neq 0$ which allows us to change the modulus of the complex number $\tau(y)$.

First let $z := \tau(c) = c_j + ic_i$ and map the center $\lambda\Lambda^{-1}c$ of the ball $E_\gamma(c)$ to

$$z' := \tau(\lambda\Lambda^{-1}c) = \lambda(\frac{c_j}{\lambda_2} + i\frac{c_i}{\lambda_1}).$$

One obtains for the angles $\varphi = \measuredangle(c, \lambda\Lambda^{-1}c)$ and $\bar{\varphi} = \measuredangle(w_1, \lambda\Lambda^{-1}c)$, cf. Fig. 5,

$$\sin\varphi = \frac{\|z-z'\|}{\|z'\|} =: \xi, \qquad \sin\bar{\varphi} = \frac{\gamma\|z-z'\|}{\|z'\|} = \gamma\xi. \tag{42}$$

By rotating z counterclockwise by φ and clockwise by $\bar{\varphi}$ one obtains

$$\tilde{w}_1 = ze^{i(\varphi-\bar{\varphi})} \tag{43}$$

with $\lambda(\tau^{-1}(\tilde{w}_1)) = \lambda(w_1)$. Combining (42) and (43) results in

$$\tilde{w}_1 = ze^{i(\arcsin\xi - \arcsin(\gamma\xi))} = ze^{i\arcsin\left(\xi(\sqrt{1-\gamma^2\xi^2}-\gamma\sqrt{1-\xi^2})\right)} = ze^{i\arcsin\rho^+}$$

with $\rho^+ = \xi\left(\sqrt{1-\gamma^2\xi^2} - \gamma\sqrt{1-\xi^2}\right)$. If $e_i \notin C_\gamma(c)$, then we obtain similarly the minimum point

$$\tilde{w}_2 = ze^{i(\varphi+\bar{\varphi})}$$

with $\lambda(\tau^{-1}(\tilde{w}_2)) = \lambda(w_2)$ and

$$\tilde{w}_2 = ze^{i(\arcsin\xi + \arcsin(\gamma\xi))} = ze^{i\arcsin\rho^-} \qquad \text{with} \qquad \rho^- = \xi\left(\sqrt{1-\gamma^2\xi^2} + \gamma\sqrt{1-\xi^2}\right).$$

Evaluating the Rayleigh quotients $\lambda(\tau^{-1}(ze^{i\arcsin\rho^{\pm}}))$ yields

$$\left[\lambda(\tau^{-1}(ze^{i\arcsin\rho^{\pm}}))\right]^{-1} = \frac{1}{\lambda_i}\left(c_i\sqrt{1-(\rho^{\pm})^2} + c_j\rho^{\pm}\right)^2 + \frac{1}{\lambda_{i+1}}\left(c_j\sqrt{1-(\rho^{\pm})^2} - c_i\rho^{\pm}\right)^2$$

which results in (39).

Finally, we have to show (41). The normalization $\|c\| = 1$ together with $\lambda(c) = \lambda$ results for the components of the componentwise non-negative vector $(c_i, c_j)^T$ in

$$c_i^2 = \frac{\lambda_i(\lambda_j-\lambda)}{\lambda(\lambda_j-\lambda_i)}, \qquad c_j^2 = \frac{\lambda_j(\lambda-\lambda_i)}{\lambda(\lambda_j-\lambda_i)}. \tag{44}$$

Using (44), elementary calculations show that

$$\xi^2 = \frac{\|c-\lambda\Lambda^{-1}c\|^2}{\|\lambda\Lambda^{-1}c\|^2} = \frac{(\lambda-\lambda_i)(\lambda_{i+1}-\lambda)}{\lambda(\lambda_i+\lambda_{i+1}-\lambda)}.$$

□

The following theorem provides an interesting link between Thm. 4.4 and Thm. 1 in [8].

Theorem 4.5. *On the assumptions of Thm. 4.4 and using* $\lambda^\pm$ *as abbreviation for (39) it holds that*

$$\frac{\lambda^\pm-\lambda_i}{\lambda_j-\lambda^\pm} = q_\pm^2 \frac{\lambda-\lambda_i}{\lambda_j-\lambda} \tag{45}$$

with q_+ *(*q_-*) being associated with* λ^+ *by (37) (*λ^- *by (38)). The convergence factors* $q_\pm$ *fulfill*

$$q_\pm = \frac{\lambda_i}{\lambda_j} \pm \gamma\left(1-\frac{\lambda_i}{\lambda_j}\right)\sqrt{c_i^2+q_\pm^2\, c_j^2} \tag{46}$$

where negative q_- *is substituted by 0 and simultaneously* $\lambda^- < \lambda_i$ *is set to* λ_i*.*

Proof. Let $\alpha = \measuredangle(z,e_i)$, $\alpha' = \measuredangle(z',e_i)$ and $\alpha_k = \measuredangle(w_k,e_i)$, $k=1,2$, be the angles between each z, z', w_1 and w_2 and the imaginary axis. As $\varphi = \alpha-\alpha'$ and $\bar\varphi = \alpha_1-\alpha'$ one has by (42)

$$\gamma = \frac{\sin(\alpha_1-\alpha')}{\sin(\alpha-\alpha')} = \frac{\sin\alpha_1 - \tan\alpha'\cos\alpha_1}{\sin\alpha_1-\tan\alpha'\cos\alpha}. \tag{47}$$

With $\sin\alpha = c_j$ and $\cos\alpha = c_i$ this results in

$$\sin\alpha_1 - \tan\alpha'\cos\alpha_1 = \gamma(c_j - c_i\tan\alpha').$$

By using $\tan\alpha' = (\lambda_i/\lambda_j)\tan\alpha$ to eliminate $\tan\alpha'$ in (47) one is led to

$$q_+ := \frac{\tan\alpha_1}{\tan\alpha} = \frac{\lambda_i}{\lambda_j} + \frac{\gamma}{\cos\alpha_1}\left(\frac{c_j}{\tan\alpha} - c_i\frac{\lambda_i}{\lambda_j}\right).$$

The latter equation can be reformulated into an equation for q_+ by using $\tan\alpha = c_j/c_i$. This yields

$$q_+ = \frac{\lambda_i}{\lambda_j} + \gamma\left(1-\frac{\lambda_i}{\lambda_j}\right)\sqrt{c_i^2+q_+^2\, c_j^2}.$$

Similarly one can derive for α_2 (instead of α_1) an equation in $q_- = \tan\alpha_2/\tan\alpha$

$$q_- = \frac{\lambda_i}{\lambda_j} - \gamma\left(1-\frac{\lambda_i}{\lambda_j}\right)\sqrt{c_i^2+q_-^2\, c_j^2}.$$

With $\cos\alpha = c_i$, $\sin\alpha = c_j$ being determined by (44) and

$$\cos\alpha_1 = (w_1)_i = \left(\frac{\lambda_i(\lambda_j-\lambda^+)}{\lambda^+(\lambda_j-\lambda_i)}\right)^{1/2}, \qquad \sin\alpha_1 = (w_1)_j = \left(\frac{\lambda_j(\lambda^+-\lambda_i)}{\lambda^+(\lambda_j-\lambda_i)}\right)^{1/2},$$

as well as the corresponding expressions for α_2, one immediately derives

$$\frac{\lambda^\pm-\lambda_i}{\lambda_j-\lambda^\pm} = q_\pm^2\frac{\lambda-\lambda_i}{\lambda_j-\lambda}.$$

□

The explicit solution of (46) for $q_\pm$ yields

$$q_\pm = \frac{\lambda_i\lambda \pm \gamma(1-\lambda_i/\lambda_j)\sqrt{\lambda_i\lambda_j\lambda(\lambda_i+\lambda_j-\lambda)+\lambda_i\lambda_j\gamma^2(\lambda_j-\lambda)(\lambda_i-\lambda)}}{\lambda_j\lambda-\gamma^2(\lambda-\lambda_i)(\lambda_j-\lambda_i)} \tag{48}$$

wherein $q_- < 0$ is set to 0. As $q_+[\lambda]$ for $\lambda \in [\lambda_i, \lambda_j]$ takes its maximum in $\lambda = \lambda_i$ (see Theorem 1 in [8]) one obtains as a λ-independent convergence factor

$$\hat{q}_+ := q_+[\lambda = \lambda_i] = \frac{\lambda_i}{\lambda_j} + \gamma\left(1 - \frac{\lambda_i}{\lambda_j}\right)$$

with

$$\frac{\lambda^+ - \lambda_i}{\lambda_j - \lambda^+} \le \hat{q}_+ \frac{\lambda - \lambda_i}{\lambda_j - \lambda}$$

for all $\lambda \in (\lambda_, \lambda_j)$. Similarly, one can derive for the maximum of $q_-[\lambda]$ which is taken in $\lambda = \lambda_j$

$$\hat{q}_- := q_-[\lambda = \lambda_j] = \frac{\lambda_i}{\lambda_j + \gamma(\lambda_j - \lambda_i)}. \tag{49}$$

In general it holds that

$$q_+ = \frac{\lambda_i}{\lambda_j} + \gamma\left(1 - \frac{\lambda_i}{\lambda_j}\right) - \varepsilon_+, \qquad q_- = \frac{\lambda_i}{\lambda_j + \gamma(\lambda_j - \lambda_i)} - \varepsilon_-$$

with $0 \le \varepsilon_+ = \mathcal{O}(\lambda - \lambda_i)$ and $0 \le \varepsilon_- = \mathcal{O}(\lambda_j - \lambda)$.

5. Convergence estimates

In this section sharp estimates are presented on the fastest possible convergence of (7) or, equivalently, on the solution of the nested optimization problem (10). The following central proof combines the results from Sec. 3 on the inner optimization problem (the choice of the preconditioner) with those on the outer problem (the choice from the level set) which has been treated in Sec. 4. Theorem 5.1 provides estimates for the different combinations of best/poorest preconditioning and best/poorest choice from the level set. These combinations are:

1. Optimal preconditioning and optimal choice from the level set results in the smallest attainable Rayleigh quotient
$$\min_{c\in L(\lambda)} \min_{d\in E_\gamma(c)} \lambda(d), \tag{50}$$
which is the case of the *fastest possible convergence.* An explicit (sharp) expression for (50) in terms of λ^- by (39) is given in Thm. 5.1.
2. Poorest preconditioning but optimal choice from the level set results in
$$\min_{c\in L(\lambda)} \max_{d\in E_\gamma(c)} \lambda(d). \tag{51}$$
3. Optimal preconditioning but poorest choice from the level set leads to
$$\max_{c\in L(\lambda)} \min_{d\in E_\gamma(c)} \lambda(d). \tag{52}$$

Note that the remaining case $\max_{c\in L(\lambda)} \max_{d\in E_\gamma(c)} \lambda(d)$ has already been treated in [10, 11]. Thm. 5.1 is formulated with respect to the c-basis introduced in Sec. 2.2, but all estimates hold with respect to the initial basis in the same manner.

Theorem 5.1. *Let both $\gamma \in [0,1)$ and $\lambda \in (\lambda_1, \lambda_n)$ be given. Then the following convergence estimates for (7) in terms of the reformulation (50)–(52) hold:*

1. *If $e_1 \in C_\gamma(c)$, then (7) for the best choice of a preconditioner can terminate in a single step within an eigenvector corresponding to the smallest eigenvalue λ_1, see Lemma 3.1.*
   If $e_1 \notin C_\gamma(c)$ and $\lambda \in (\lambda_1, \lambda_2)$, then the minimum (50) reads
   $$\lambda^-(\lambda_1, \lambda_n, \lambda, \gamma) = \min_{c \in L(\lambda)} \min_{d \in E_\gamma(c)} \lambda(d) = \min_{d \in E_\gamma(c_{1,n})} \lambda(d), \tag{53}$$
   with $\lambda^-(\lambda_1, \lambda_n, \lambda, \gamma)$ being defined by (39). In (53) $c_{1,n}$ is a vector of the form
   $$c_{1,n} := (c_1, 0, \ldots, 0, c_n)^T \in L(\lambda),$$
   i.e., the minimum (53) is attained in a 2D subspace spanned by the eigenvectors corresponding to λ_1 and λ_n.
   If $\lambda \in [\lambda_i, \lambda_{i+1})$, $i > 1$, then
   $$\min_{c \in L(\lambda)} \min_{d \in E_\gamma(c)} \lambda(d) \leq \lambda^-(\lambda_1, \lambda_n, \lambda, \gamma). \tag{54}$$
2. *Poorest preconditioning within the vector $c_{1,n} \in L(\lambda)$ of fastest convergence (of case 1.) leads to*
   $$\min_{c \in L(\lambda)} \max_{d \in E_\gamma(c)} \lambda(d) \leq \lambda^+(\lambda_1, \lambda_n, \lambda, \gamma) = \max_{d \in E_\gamma(c_{1,n})} \lambda(d) \tag{55}$$
   with $\lambda^+(\lambda_1, \lambda_n, \lambda, \gamma)$ being defined by (39).
3. *Let $e_1 \notin C_\gamma(c)$. Then optimal preconditioning within the vector $c_{i,i+1} \in L(\lambda)$ of slowest convergence (see Thm. 1.1 in [11]) results in the Rayleigh quotient*
   $$\min_{d \in E_\gamma(c_{i,i+1})} \lambda(d) \leq \lambda^-(\lambda_i, \lambda_{i+1}, \lambda, \gamma). \tag{56}$$

Proof. To show (53), the idea is to compare the decrease of the Rayleigh quotient along the curves of extremum points as derived in Thm. 3.4. For a given $y \in L(\lambda)$ such a curve, by (28), has the form
$$S(y) := \frac{(\alpha I + \Lambda)^{-1} y}{\|(\alpha I + \Lambda)^{-1} y\|}, \qquad \alpha \in (\alpha_{\min}(y), \infty)$$
for certain $\alpha_{\min}(y) \geq -\lambda_1$. On the one hand, we take the curve $S(c_{1,n})$, $c_{1,n} \in L(\lambda)$, which starts on the level set $L(\lambda)$ for $\alpha \to \infty$ and runs along all extremum points of $C_\gamma(c_{1,n})$ for all $\gamma \in [0,1]$. We follow this curve until the (normalized) minimum point on $C_\gamma(c_{1,n})$ is reached. On the other hand, we take a second curve $S(c)$ for arbitrary $c \in L(\lambda)$, $c \neq c_{1,n}$. Once again, $S(c)$ starts on the level set $L(\lambda)$. Our aim is to derive (53) by proving that the Rayleigh quotient along $S(c_{1,n})$ decreases faster than on $S(c)$.

First note that by Lemma 4.2 the opening angle φ_γ of C_γ takes its maximum on $L(\lambda)$ in $c_{1,n}$, i.e.
$$\varphi_\gamma(c_{1,n}) \geq \varphi_\gamma(c), \qquad \forall \gamma \in [0,1], \quad \forall c \in L(\lambda). \tag{57}$$

We parametrize each $S(c)$ and $S(c_{1,n})$ in an angle variable φ in the following manner: The curve $S(c)$ starts at c for $\varphi = 0$ reaches the axis $\lambda \Lambda^{-1} c$ of $C_\gamma(c)$ for $\varphi_1(c)$ and ends in the minimum point of $C_\gamma(c)$ for $\varphi_\gamma(c) + \varphi_1(c)$. In the same way the curve $S(c_{1,n})$ is parametrized in φ. Thus $S(c_{1,n})$ starts at $c_{1,n}$ for $\varphi = 0$ and ends in the minimum point of $C_\gamma(c_{1,n})$ for $\varphi_\gamma(c_{1,n}) + \varphi_1(c_{1,n})$. For the angles in the minimum points (57) yields
$$\varphi_\gamma(c_{1,n}) + \varphi_1(c_{1,n}) \geq \varphi_\gamma(c) + \varphi_1(c). \tag{58}$$

The corresponding Rayleigh quotients on these curves parametrized in φ are denoted by $\lambda(c,\varphi)$ and $\lambda(c_{1,n},\varphi)$. Then for any pair of angles $\tilde{\varphi}$ and $\tilde{\varphi}_{1,n}$ with

$$\lambda(c,\tilde{\varphi}) = \lambda(c_{1,n},\tilde{\varphi}_{1,n})$$

by Lemma 2.6 in [11] and Lemma 4.3 together with Lemma 4.1 it holds that

$$\left|\frac{d\lambda(c,\varphi)}{d\varphi}\Big|_{\varphi=\tilde{\varphi}}\right| \leq \left|\frac{d\lambda(c_{1,n},\varphi)}{d\varphi}\Big|_{\varphi=\tilde{\varphi}_{1,n}}\right|. \tag{59}$$

Inequality (59) proves a locally faster decrease of the Rayleigh quotient along the curve $S(c_{1,n})$.

Hence $f(\varphi) := \lambda(c_{1,n},\varphi)$ and $g(\varphi) := \lambda(c,\varphi)$ are monotone decreasing, differentiable positive functions. Eq. (59) simply says that in all arguments α,β with $f(\alpha) = g(\beta)$, the (negative) derivatives fulfill

$$f'(\alpha) \leq g'(\beta).$$

Hence, because of $f(0) = g(0)$ it holds that

$$f(\xi) \leq g(\xi),$$

with ξ being the smaller angle $\xi = \varphi_\gamma(c) + \varphi_1(c)$ in (58). Monotonicity of f shows that for the larger angle $\varphi_\gamma(c_{1,n}) + \varphi_1(c_{1,n})$ it holds that

$$\lambda(c_{1,n},\varphi_\gamma(c_{1,n}) + \varphi_1(c_{1,n})) \leq \lambda(c,\varphi_\gamma(c) + \varphi_1(c)),$$

which proves faster decrease of the Rayleigh quotient along $S(c_{1,n})$ compared to $S(c)$. The value of $\lambda(c_{1,n},\varphi_\gamma(c_{1,n}) + \varphi_1(c_{1,n}))$ can be derived in the 2D invariant subspace spanned by e_1 and e_n since $S(c_{1,n}) \subseteq \operatorname{span}\{e_1,e_n\}$. The mini-dimensional analysis in Thm. 4.4 for $i = 1$ and $j = n$ proves that the minimum is given by (38), i.e., $\lambda' = \lambda^-(\lambda_1,\lambda_n,\lambda,\gamma)$.

To prove (54) we use the same construction as above. Once again we compare $S(c_{1,n})$ with $S(c)$. Inequality (54) is not necessarily sharp, as a possible bifurcation (for $c \in L(\lambda)$ with $c_1 = 0$, see Sec. 3.2) is not taken into account.

To show (55), we proceed as in the first part of the proof. Now we compare the curves of maximum points for c and $c_{1,n} \in L(\lambda)$. These curves are the initial parts of the curves $S(c)$ and $S(c_{1,n})$ considered above. We follow these curves along their parametrization in φ until the maximum points are reached. These maximum points are reached within φ equal to certain *shrinking angles* ψ_γ (see Lemma 2.3 in [11]). It holds

$$\psi_\gamma(c_{1,n}) \geq \psi_\gamma(c).$$

Along these curves (59) holds for any φ and $\varphi_{1,n}$ with $\lambda(c,\varphi) = \lambda(c_{1,n},\varphi_{1,n})$; the latter Rayleigh quotients are now associated with maximum points. Once again, (59) proves that the Rayleigh quotient decreases locally faster along $S(c_{1,n})$. In analogy to the derivation above we obtain

$$\lambda(c,\psi_\gamma(c)) \geq \lambda(c_{1,n},\psi_\gamma(c_{1,n})),$$

which proves a globally faster decrease of the Rayleigh quotient on $S(c_{1,n})$. The Rayleigh quotient $\lambda^+(\lambda_1,\lambda_n,\lambda,\gamma)$ results from applying the mini-dimensional analysis to the 2D space $\operatorname{span}\{e_1,e_n\}$, see Sec. 4.2.

Finally, to show (56), we proceed similarly to the first case but compare the Rayleigh quotients along the extremum curves associated with $c,\ c_{i,i+1} \in L(\lambda)$. The Rayleigh quotient $\lambda^-(\lambda_i,\lambda_{i+1},\lambda,\gamma)$ is only an upper bound in (56), since by construction all minima are constrained to $\operatorname{span}\{e_i, e_{i+1}\}$. Therefore a possible bifurcation of the minimum curve is disregarded and λ^- is larger than the minimal Rayleigh quotient on E_γ. □

The convergence estimates of Thm. 5.1 are difficult to grasp due to the complex nature of $\lambda^\pm(\lambda_i, \lambda_j, \lambda, \gamma)$. By Thm. 4.5 the next corollary follows immediately. Estimate (60) can be applied recursively as $\hat{q}_-$ does not depend on $\lambda(x)$.

Corollary 5.2. *Assume an optimal choice of $x \in L(\lambda)$ with $\lambda < \lambda_2$ and optimal preconditioning in the sense of Sec. 3. Then the fastest possible decrease of $\lambda(x')$ with x' by (7) toward the smallest eigenvalue λ_1 is bounded from above by*

$$\frac{\lambda(x') - \lambda_1}{\lambda_n - \lambda(x')} \leq \hat{q}_-^2 \frac{\lambda(x) - \lambda_1}{\lambda_n - \lambda(x)} \tag{60}$$

with the convergence factor

$$\hat{q}_- = \frac{\lambda_1}{\lambda_n + \gamma(\lambda_n - \lambda_1)}.$$

Proof. By Thm. 5.1 fastest convergence with respect to the level set $L(\lambda)$ is taken in the 2D subspace spanned by the eigenvectors to λ_1 and λ_n. Thus (60) follows from (49) for $i = 1$ and $j = n$; see also Thm. 4.6 in [12]. □

In the following the convergence estimates of Thm. 5.1 are illustrated for a low-dimensional model problem with the eigenvalues $(\lambda_1, \ldots, \lambda_6) = (2, 5, 8, 10, 13, 17)$, i.e., the first eigenvalues of Laplace operator on $[0, \pi]^2$. This is the same example which has already been used in Fig. 1 in [11].

In Fig. 6 the quotients

$$\Phi^\pm_{i,j}(\lambda, \gamma) := \frac{\lambda^\pm(\lambda_i, \lambda_j, \lambda, \gamma) - \lambda_i}{\lambda - \lambda_i} \leq 1, \tag{61}$$

which measure the *relative decrease of $\lambda^\pm(\lambda_i, \lambda_j, \lambda, \gamma)$ toward the next smaller eigenvalue λ_i*, are drawn for $\lambda \in [2, 17]$. The different curves are each plotted for $\gamma = 0, 0.1, \ldots, 1.0$.

The convergence factors (61) measure the relative decrease of the error of the eigenvalue approximations; they guarantee the convergence of the iterates to an eigenpair as the ratios are bounded from above by 1. In the interval $[\lambda_i, \lambda_{i+1}]$ the convergence factor $\Phi^\pm_{i,i+1}$ is a function of λ and γ.

First, in the upper part of Fig. 6 the curves $\Phi^+_{i,i+1}(\lambda, \gamma)$ are shown; see also [11] for an explanation of the fan-like structure of these bounds. The discontinuity from $\Phi^+_{i,i+1}$ to $\Phi^+_{i-1,i}$ in $\lambda = \lambda_i$ reflects that poorest decrease of the Rayleigh quotient corresponds to an early breakdown of the iteration in an eigenvector corresponding to λ_i. The quotients $\Phi^-_{1,n}(\lambda, \gamma)$ correspond to the fastest convergence, i.e., the fastest decrease of the Rayleigh quotient. The assumption $e_1 \notin C_\gamma(c)$ in Thm. 5.1 is only made to avoid tiresome case distinctions. Whenever for a certain $\lambda^* \in [\lambda_1, \lambda_n]$ it holds that $\lambda^-_{1,n}(\lambda^*, \gamma^*) = \lambda_1$, then $e_1 \in C_\gamma(c)$ for all $\gamma \geq \gamma^*$. Hence, what is actually drawn in Fig. 6 is

$$\tilde{\Phi}^-_{1,n}(\lambda, \gamma) := \min_{\tilde{\gamma} \leq \gamma} \Phi^-_{1,n}(\lambda, \tilde{\gamma}). \tag{62}$$

Finally, in the lower part of Fig. 6 the remaining curves are illustrated. They correspond to the best choice in $L(\lambda)$ together with poorest preconditioning (case $\Phi^+_{1,n}$, see dotted lines) and poorest choice from $L(\lambda)$ together with best preconditioning in $\operatorname{span}\{e_i, e_{i+1}\}$, i.e., the case $\Phi^-_{i,i+1}$ as drawn by the dashed lines.

Fig. 6 exhibits a wide range between fastest and slowest convergence. The one extreme is stationarity ($\Phi \to 1$) and the other extreme is one-step convergence ($\Phi \to 0$). Note that the estimates on slowest convergence in $[\lambda_i, \lambda_{i+1}]$ do not depend on the largest eigenvalue λ_n (aside

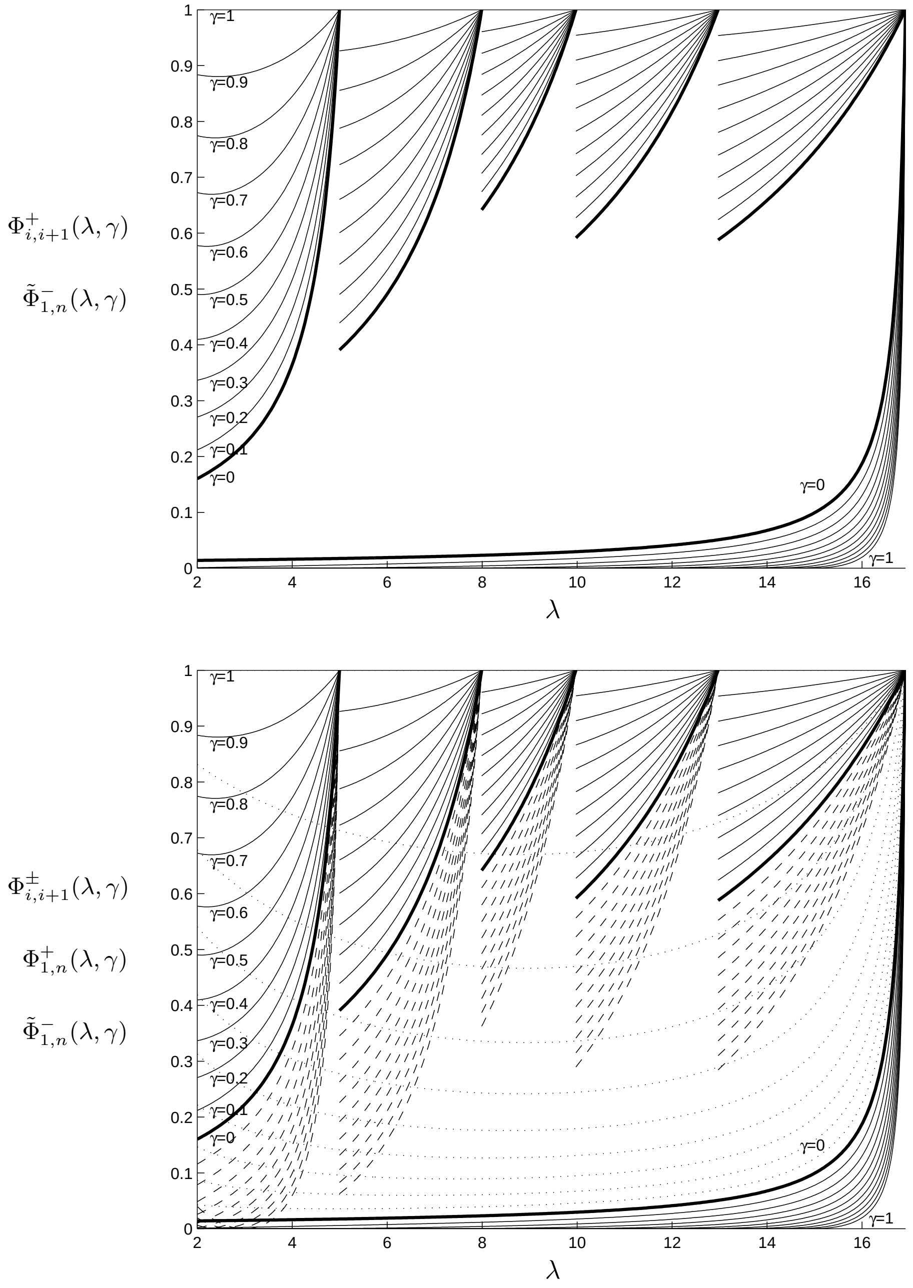

Figure 6. *Bounds on the fastest and slowest convergence for the model problem* $\Lambda = \mathrm{diag}(2,5,8,10,13,17)$ *with* $\gamma = 0, 0.1, \ldots, 1$.

Upper figure: Bounds $\Phi^{+}_{i,i+1}$ *(slowest convergence) and* $\tilde{\Phi}^{-}_{1,n}$ *(fastest convergence) by (61) and (62).*

Lower figure: Additionally drawn are $\Phi^{-}_{i,i+1}$ *(dashed) and* $\Phi^{+}_{1,n}$ *(dotted).*

from $i+1=n$), but that the quotient $\Phi^{\pm}_{1,n}$ does so. Hence, whenever λ_n increases, the corridor between slowest and fastest convergence widens, making even faster convergence possible. Let us now determine the particular λ^* in $\operatorname{span}\{e_1, e_n\}$, below which one-step convergence to λ_1 is possible. Condition (18) in $\operatorname{span}\{e_1, e_n\}$ leads to

$$\lambda^* = \lambda_n \left(1 + \frac{\lambda_1}{\gamma^2(\lambda_n - \lambda_1)}\right)^{-1}. \tag{63}$$

so that

$$\lambda^* = \mathcal{O}(h^{-2})$$

for the discrete Laplacian Δ_h. There is also a critical bound γ^* so that for $\gamma < \gamma^*$ the eigenvector e_1 is never contained in $C_\gamma(c)$. Setting $\lambda^* = \lambda_1$ in (63) and solving for γ results in

$$\gamma^* = \frac{\lambda_1}{\lambda_n - \lambda_1} = \mathcal{O}(h^2)$$

for Δ_h. Hence, one-step convergence is impossible if B approximates A very accurately.

The bold curves in Fig. 6 are associated with $\gamma = 0$ or inverse iteration. Not surprisingly, (7) may converge faster than inverse iteration since $\min \lambda(E_\gamma(c)) < \lambda(\lambda\Lambda^{-1}c)$. The upper bold curves correspond to $c_{i,i+1}$, $i = 1, \dots, 4$, whereas the lower bold curve corresponds to $c_{1,n}$.

6. Conclusion

Sharp convergence estimates on the fastest convergence have been derived for a basic preconditioned eigensolver. This analysis is based on a geometrical approach which has proved very useful for understanding the extremal convergence behavior. The key point of this geometrical approach is that the set of possible iterates, which is generated by all admissible preconditioners, is a ball with respect to the A-geometry. In the light of the present analysis several practical questions (which are not treated here) appear very clearly. Among others the following questions are provoked: How to practically find/construct a preconditioner which leads to fast convergence? How to generate an appropriate initial iteration vector?

Within the framework of a geometrical interpretation it is immediately clear that exact preconditioning, i.e., $B = A$, is not the optimal choice for solving an eigenvalue problem. Instead, optimal preconditioning, under the condition of Lemma 3.1, allows even one-step convergence to an eigenpair.

The convergence estimates which have been derived here are not only of theoretical value. They can explain that sometimes (especially in the first steps of an iteration), the scheme (7) may converge much more rapidly than suggested by the worst case estimates presented in [10, 11]. Implicitly, it is shown that a lot of space is left for accelerating the basic preconditioning eigensolver (7); a development in this sense is the practically important *locally optimal preconditioned conjugate gradient* iteration [7].

REFERENCES

1. 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.
2. E.G. D'yakonov. Iteration methods in eigenvalue problems. *Math. Notes*, 34:945–953, 1983.
3. E.G. D'yakonov. *Optimization in solving elliptic problems.* CRC Press, Boca Raton, Florida, 1996.

4. 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.
5. A.V. Knyazev. Private communication. 2000.
6. A.V. Knyazev. Preconditioned eigensolvers—an oxymoron? *Electron. Trans. Numer. Anal.*, 7:104–123, 1998.
7. A.V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM J. Sci. Comp.*, 23:517–541, 2001.
8. A.V. Knyazev and K. Neymeyr. A geometric theory for preconditioned inverse iteration. III: A short and sharp convergence estimate for generalized eigenvalue problems. *Linear Algebra Appl.*, 358:95–114, 2003.
9. A.V. Knyazev and K. Neymeyr. Efficient solution of symmetric eigenvalue problems using multigrid preconditioners in the locally optimal block conjugate gradient method. *Electron. Trans. Numer. Anal.*, 15:38–55, 2003.
10. K. Neymeyr. A geometric theory for preconditioned inverse iteration. I: Extrema of the Rayleigh quotient. *Linear Algebra Appl.*, 322:61–85, 2001.
11. K. Neymeyr. A geometric theory for preconditioned inverse iteration. II: Convergence estimates. *Linear Algebra Appl.*, 322:87–104, 2001.
12. K. Neymeyr. A hierarchy of preconditioned eigensolvers for elliptic differential operators. Habilitation thesis, Universität Tübingen, Germany, 2001.
13. K. Neymeyr. A geometric theory for preconditioned inverse iteration applied to a subspace. *Math. Comp.*, 71:197–216, 2002.
14. B.A. Samokish. The steepest descent method for an eigenvalue problem with semi-bounded operators. *Izv. Vyssh. Uchebn. Zaved. Mat.*, 5:105–114, 1958. (In Russian).