

## GRADIENT FLOW APPROACH TO GEOMETRIC CONVERGENCE ANALYSIS OF PRECONDITIONED EIGENSOLVERS \*

ANDREW V. KNYAZEV<sup>†</sup> AND KLAUS NEYMEYR<sup>‡</sup>

**Abstract.** Preconditioned eigenvalue solvers (eigsolvers) are gaining popularity, but their convergence theory remains sparse and complex. We consider the simplest preconditioned eigensolver—the gradient iterative method with a fixed step size—for symmetric generalized eigenvalue problems, where we use the gradient of the Rayleigh quotient as an optimization direction. A sharp convergence rate bound for this method has been obtained in 2001–2003. It still remains the only known such bound for any of the methods in this class. While the bound is short and simple, its proof is not. We extend the bound to Hermitian matrices in the complex space and present a new self-contained and significantly shorter proof using novel geometric ideas.

**Key words.** iterative method; continuation method; preconditioning; preconditioner; eigenvalue; eigenvector; Rayleigh quotient; gradient iteration; convergence theory; spectral equivalence

**AMS subject classifications.** 49M37 65F15 65K10 65N25

(Place for Digital Object Identifier, to get an idea of the final spacing.)

**1. Introduction.** We consider a generalized eigenvalue problem (eigenproblem) for a linear pencil  $B - \mu A$  with symmetric (Hermitian in the complex case) matrices  $A$  and  $B$  with positive definite  $A$ . The eigenvalues  $\mu_i$  are enumerated in decreasing order  $\mu_1 \geq \dots \geq \mu_{\min}$  and the  $x_i$  denote the corresponding eigenvectors. The largest value of the Rayleigh quotient  $\mu(x) = (x, Bx)/(x, Ax)$ , where  $(\cdot, \cdot)$  denotes the standard scalar product, is the largest eigenvalue  $\mu_1$ . It can be approximated iteratively by maximizing the Rayleigh quotient in the direction of its gradient, which is proportional to  $(B - \mu(x)A)x$ . Preconditioning is used to accelerate the convergence; see, e.g., [2, 6, 4, 7, 8] and the references therein. Here we consider the simplest preconditioned eigenvalue solver (eigsolver)—the gradient iterative method with an explicit formula for the step size, cf. [2], one step of which is described by

$$(1.1) \quad x' = x + \frac{1}{\mu(x) - \mu_{\min}} T(Bx - \mu(x)Ax), \quad \mu(x) = \frac{(x, Bx)}{(x, Ax)}.$$

The symmetric (Hermitian in the complex case) positive definite matrix  $T$  in (1.1) is called the *preconditioner*. Since  $A$  and  $T$  are both positive definite, we assume that

$$(1.2) \quad (1 - \gamma)(z, T^{-1}z) \leq (z, Az) \leq (1 + \gamma)(z, T^{-1}z), \quad \forall z, \text{ for a given } \gamma \in [0, 1).$$

The following result is proved in [8, 9, 10] in the real space.

**THEOREM 1.1.** *If  $\mu_{i+1} < \mu(x) \leq \mu_i$  then  $\mu(x') \geq \mu(x)$  and*

$$(1.3) \quad \frac{\mu_i - \mu(x')}{\mu(x') - \mu_{i+1}} \leq \sigma^2 \frac{\mu_i - \mu(x)}{\mu(x) - \mu_{i+1}}, \quad \sigma = 1 - (1 - \gamma) \frac{\mu_i - \mu_{i+1}}{\mu_i - \mu_{\min}}.$$

*The convergence factor  $\sigma$  cannot be improved with the chosen terms and assumptions.*

---

\*Received by the editors February 8, 2009; accepted for publication (in revised form) ??????, 2009; published electronically ??????, 2009. Preliminary version <http://arxiv.org/abs/0801.3099>

<sup>†</sup> Department of Mathematical and Statistical Sciences, University Colorado Denver, P.O. Box 173364, Campus Box 170, Denver, CO 80217-3364 (andrew.knyazev at ucdenver.edu, <http://math.ucdenver.edu/~aknyazev/>). Supported by the NSF-DMS 0612751.

<sup>‡</sup>Universität Rostock, Institut für Mathematik, Universitätsplatz 1, 18055 Rostock, Germany (klaus.neymeyr at mathematik.uni-rostock.de, <http://cat.math.uni-rostock.de/~neymeyr/>).

Compared to other known non-asymptotic convergence rate bounds for similar preconditioned eigensolvers, e.g., [2, 6, 7, 1], the advantages of (1.3) are in its sharpness and elegance. Method (1.1) is the easiest preconditioned eigensolver, but (1.3) still remains the only known sharp bound in these terms for any of preconditioned eigensolvers. While bound (1.3) is short and simple, its proof in [8] is quite the opposite. It covers only the real case and is not self-contained—in addition it requires most of the material from [9, 10]. Here we extend the bound to Hermitian matrices and give a new much shorter and self-contained proof of Theorem 1.1, which is a great qualitative improvement compared to that of [8, 9, 10]. The new proof is not yet as elementary as we would like it to be; however, it is easy enough to hope that a similar approach might be applicable in future work on preconditioned eigensolvers.

Our new proof is based on novel techniques combined with some old ideas of [5, 9, 10]. We demonstrate that, for a given initial eigenvector approximation  $x$ , the next iterative approximation  $x'$  described by (1.1) belongs to a cone if we apply any preconditioner satisfying (1.2). We analyse a corresponding continuation gradient method involving the gradient flow of the Rayleigh quotient and show that the smallest gradient norm (evidently leading to the slowest convergence) of the continuation method is reached when the initial vector belongs to a subspace spanned by two specific eigenvectors, namely  $x_i$  and  $x_{i+1}$ . This is done by showing that Temple's inequality, which provides a lower bound for the norm of the gradient  $\nabla\mu(x)$ , is sharp only in  $\text{span}\{x_i, x_{i+1}\}$ . Next, we extend by integration the result for the continuation gradient method to our actual fixed step gradient method to conclude that the point on the cone, which corresponds to the poorest convergence and thus gives the guaranteed convergence rate bound, belongs to the same two-dimensional invariant subspace  $\text{span}\{x_i, x_{i+1}\}$ . This reduces the convergence analysis to a two-dimensional case for shifted inverse iterations, where the sharp convergence rate bound is established.

## 2. The proof of Theorem 1.1.

We start with several simplifications:

**THEOREM 2.1.** *We can assume that  $\gamma > 0$ ,  $A = I$ ,  $B > 0$  is diagonal, eigenvalues are simple,  $\mu(x) < \mu_i$ , and  $\mu(x') < \mu_i$  in Theorem 1.1 without loss of generality.*

*Proof.* First, we observe that method (1.1) and bound (1.3) are evidently both invariant with respect to a real shift  $s$  if we replace the matrix  $B$  with  $B + sA$ , so without loss of generality we need only consider the case  $\mu_{\min} = 0$  which makes  $B \geq 0$ . Second, by changing the basis from coordinate vectors to the eigenvectors of  $A^{-1}B$  we can make  $B$  diagonal and  $A = I$ . Third, having  $\mu(x') \geq \mu(x)$  if  $\mu(x) = \mu_i$  or  $\mu(x') \geq \mu_i$ , or both, bound (1.3) becomes trivial. The assumption  $\gamma > 0$  is a bit more delicate. The vector  $x'$  depends continuously on the preconditioner  $T$ , so we can assume that  $\gamma > 0$  and extend the final bound to the case  $\gamma = 0$  by continuity.

Finally, we again use continuity to explain why we can assume that all eigenvalues (in fact, we only need  $\mu_i$  and  $\mu_{i+1}$ ) are simple and make  $\mu_{\min} > 0$  and thus  $B > 0$  without changing anything. Let us list all  $B$ -dependent terms, in addition to all participating eigenvalues, in method (2.1):  $\mu(x)$  and  $x'$ ; and in bound (1.3):  $\mu(x)$  and  $\mu(x')$ . All these terms depend on  $B$  continuously if  $B$  is slightly perturbed into  $B_\epsilon$  with some  $\epsilon \rightarrow 0$ , so we increase arbitrarily small the diagonal entries of the matrix  $B$  to make all eigenvalues of  $B_\epsilon$  simple and  $\mu_{\min} > 0$ . If we prove bound (1.3) for the matrix  $B_\epsilon$  with simple positive eigenvalues, and show that the bound is sharp as  $0 < \mu_{\min} \rightarrow 0$  with  $\epsilon \rightarrow 0$ , we take the limit  $\epsilon \rightarrow 0$  and by continuity extend the result to the limit matrix  $B \geq 0$  with  $\mu_{\min} = 0$  and possibly multiple eigenvalues.  $\square$

It is convenient to rewrite (1.1)–(1.3) equivalently by Theorem 2.1 as follows<sup>1</sup>

$$(2.1) \quad \mu(x)x' = Bx - (I - T)(Bx - \mu(x)x), \quad \mu(x) = \frac{(x, Bx)}{(x, x)},$$

$$(2.2) \quad \|I - T\| \leq \gamma, \quad 0 < \gamma < 1;$$

and if  $\mu_{i+1} < \mu(x) < \mu_i$  and  $\mu(x') < \mu_i$  then  $\mu(x') \geq \mu(x)$  and

$$(2.3) \quad \frac{\mu_i - \mu(x')}{\mu(x') - \mu_{i+1}} \leq \sigma^2 \frac{\mu_i - \mu(x)}{\mu(x) - \mu_{i+1}}, \quad \sigma = 1 - (1 - \gamma) \frac{\mu_i - \mu_{i+1}}{\mu_i} = \gamma + (1 - \gamma) \frac{\mu_{i+1}}{\mu_i}$$

Now we establish the validity and sharpness of bound (2.3) assuming (2.1) and (2.2).

**THEOREM 2.2.** *Let us define<sup>2</sup>  $\phi_\gamma(x) = \arcsin(\gamma\|Bx - \mu(x)x\|/\|Bx\|)$ , then  $\phi_\gamma(x) < \pi/2$  and  $\angle\{x', Bx\} \leq \phi_\gamma(x)$ . Let  $w \neq 0$  be defined as the vector constrained by  $\angle\{w, Bx\} \leq \phi_\gamma(x)$  and with the smallest value  $\mu(w)$ . Then  $\mu(x') \geq \mu(w) > \mu(x)$ .*

*Proof.* Orthogonality  $(x, Bx - \mu(x)x) = 0$  by the Pythagorean theorem implies  $\|Bx\|^2 = \|\mu(x)x\|^2 + \|Bx - \mu(x)x\|^2$ , so  $\|Bx - \mu(x)x\| < \|Bx\|$ , since  $\mu(x) > 0$  as  $B > 0$ , and  $\sin \angle\{x, Bx\} = \sin \phi_1(x) = \|Bx - \mu(x)x\|/\|Bx\| < 1$ , where  $Bx \neq 0$  as  $B > 0$ . A ball with the radius  $\gamma\|Bx - \mu(x)x\| \geq \|I - T\|\|Bx - \mu(x)x\|$  by (2.2) centered at  $Bx$  contains  $\mu(x)x'$  by (2.1), so  $\sin \angle\{x', Bx\} \leq \gamma\|Bx - \mu(x)x\|/\|Bx\| < \gamma < 1$ .

The statement  $\mu(x') \geq \mu(w)$  follows directly from the definition of  $w$ . Now,

$$0 < \frac{(x, Bx)}{\|x\|\|Bx\|} = \cos \phi_1(x) < \cos \angle\{w, Bx\} = \frac{|(w, Bx)|}{\|w\|\|Bx\|} \leq \frac{(w, Bw)^{1/2}(x, Bx)^{1/2}}{\|w\|\|Bx\|}$$

as  $B > 0$ , so  $\sqrt{\mu(x)} < \sqrt{\mu(w)}$  and  $\mu(x) < \mu(w)$ .  $\square$

We denote by  $C_{\phi_\gamma(x)}(Bx) := \{y : \angle\{y, Bx\} \leq \phi_\gamma(x)\}$  the circular cone around  $Bx$  with the opening angle  $\phi_\gamma(x)$ . Theorem 2.2 replaces  $x'$  with the minimizer  $w$  of the Rayleigh quotient on the cone  $C_{\phi_\gamma(x)}(Bx)$  in the rest of the paper, except at the end of the proof of Theorem 2.7, where we show that bounding below the value  $\mu(w)$  instead of  $\mu(x')$  still gives the sharp estimate.

Later on, in the proof of Theorem 2.4, we use an argument that holds easily only in the real space, so we need the following last simplification.

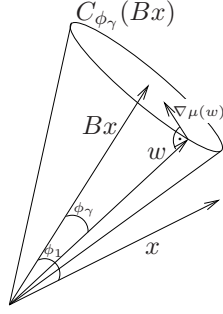
**THEOREM 2.3.** *Without loss of generality we can consider only the real case.*

*Proof.* The key observation is that for our positive diagonal matrix  $B$  the Rayleigh quotient depends evidently only on the absolute values of the vector components, i.e.,  $\mu(x) = \mu(|x|)$ , where the absolute value operation is applied component-wise. Moreover,  $\|Bx - \mu(x)x\| = \|B|x| - \mu(|x|)|x|\|$  and  $\|Bx\| = \|B|x|\|$ , so  $\phi_\gamma(x) = \phi_\gamma(|x|)$ . The cone  $C_{\phi_\gamma(x)}(Bx)$  lives in the complex space, but we also need its substitute in the real space. Let us introduce the notation  $C_{\phi_\gamma(|x|)}^R(B|x|)$  for the *real* circular cone with the opening angle  $\phi_\gamma(|x|)$  centered at the *real* vector  $B|x|$ . Next we show that in the real space we have the inclusion  $|C_{\phi_\gamma(x)}(Bx)| \subseteq C_{\phi_\gamma(|x|)}^R(B|x|)$ .

For any complex nonzero vectors  $x$  and  $y$ , we have  $|(y, Bx)| \leq (|y|, B|x|)$  by the triangle inequality, thus  $\angle\{|y|, B|x|\} \leq \angle\{y, Bx\}$ . If  $y \in C_{\phi_\gamma(x)}(Bx)$  then  $\angle\{|y|, B|x|\} \leq \angle\{y, Bx\} \leq \phi_\gamma(x) = \phi_\gamma(|x|)$ , i.e., indeed,  $|y| \in C_{\phi_\gamma(|x|)}^R(B|x|)$ , which means that  $|C_{\phi_\gamma(x)}(Bx)| \subseteq C_{\phi_\gamma(|x|)}^R(B|x|)$  as required.

<sup>1</sup>Here and below  $\|\cdot\|$  denotes the Euclidean vector norm, i.e.,  $\|x\|^2 = (x, x) = x^H x$  for a real or complex column-vector  $x$ , as well as the corresponding induced matrix norm.

<sup>2</sup>We define angles in  $[0, \pi/2]$  between vectors by  $\cos \angle\{x, y\} = |(x, y)|/(\|x\|\|y\|)$ .

FIG. 2.1. The cone  $C_{\phi_\gamma(x)}(Bx)$ .

Therefore, changing the given vector  $x$  to take its absolute value  $|x|$  and replacing the complex cone  $C_{\phi_\gamma(x)}(Bx)$  with the real cone  $C_{\phi_\gamma(|x|)}^R(B|x|)$  lead to the relations  $\min_{y \in C_{\phi_\gamma(x)}(Bx)} \mu(y) = \min_{|y| \in |C_{\phi_\gamma(x)}(Bx)|} \mu(|y|) \geq \min_{|y| \in C_{\phi_\gamma(|x|)}^R(B|x|)} \mu(|y|)$ , but does not affect the starting Rayleigh quotient  $\mu(x) = \mu(|x|)$ . This proves the theorem with the exception of the issue of whether the sharpness in the real case implies the sharpness in the complex case; see the end of the proof of Theorem 2.7.  $\square$

**THEOREM 2.4.** *We have  $w \in \partial C_{\phi_\gamma(x)}(Bx)$  and  $\exists \alpha = \alpha_\gamma(x) > -\mu_i$  such that  $(B + \alpha I)w = Bx$ . The inclusion  $x \in \text{span}\{x_i, x_{i+1}\}$  implies  $w \in \text{span}\{x_i, x_{i+1}\}$ .*

*Proof.* Assuming that  $w$  is strictly inside the cone  $C_{\phi_\gamma(x)}(Bx)$  implies that  $w$  is a point of a local minimum of the Rayleigh quotient. The Rayleigh quotient has only one local (and global) minimum,  $\mu_{\min}$ , but the possibility  $\mu(w) = \mu_{\min}$  is eliminated by Theorem 2.2, so we obtain a contradiction, thus  $w \in \partial C_{\phi_\gamma(x)}(Bx)$ .

The necessary condition for a local minimum of a smooth real-valued function on a smooth surface in a real vector space is that the gradient of the function is orthogonal to the surface at the point of the minimum and directed inwards. In our case,  $C_{\phi_\gamma(x)}(Bx)$  is a circular cone with the axis  $Bx$  and the gradient  $\nabla\mu(w)$  is positively proportional to  $Bw - \mu(w)w$ ; see Figure 2.1. We first scale the vector  $w$  such that  $(Bx - w, w) = 0$  so that the vector  $Bx - w$  is an inward normal vector for  $\partial C_{\phi_\gamma(x)}(Bx)$  at the point  $w$ . This inward normal vector must be positively proportional to the gradient,  $\beta(Bx - w) = Bw - \mu(w)w$  with  $\beta > 0$ , which gives  $(B + \alpha I)w = \beta Bx$ , where  $\alpha = \beta - \mu(w) > -\mu(w) > -\mu_i$ . Here  $\beta \neq 0$  as otherwise  $w$  would be an eigenvector, but  $\mu(x) < \mu(w) < \mu(x')$  by Theorem 2.2, where by assumptions  $\mu_{i+1} < \mu(x)$ , while  $\mu(x') < \mu_i$  by Theorem 2.1, which gives a contradiction. As the scaling of the minimizer is irrelevant, we denote  $w/\beta$  here by  $w$  with a slight local notation abuse.

Finally, since  $(B + \alpha I)w = Bx$ , inclusion  $x \in \text{span}\{x_i, x_{i+1}\}$  gives either the required inclusion  $w \in \text{span}\{x_i, x_{i+1}\}$  or  $w \in \text{span}\{x_i, x_{i+1}, x_j\}$  with  $\alpha = -\mu_j$  for some  $j \neq i$  and  $j \neq i + 1$ . We now show that the latter leads to a contradiction. We have just proved that  $\alpha > -\mu_i$ , thus  $j > i + 1$ . Let  $x = c_i x_i + c_{i+1} x_{i+1}$ , where we notice that  $c_i \neq 0$  and  $c_{i+1} \neq 0$  since  $x$  is not an eigenvector. Then we obtain  $w = a_i c_i x_i + a_{i+1} c_{i+1} x_{i+1} + c_j x_j$  where  $(B - \mu_j)w = Bx$ , therefore  $a_k = \mu_k / (\mu_k - \mu_j)$ ,  $k = i, i + 1$ . Since all eigenvalues are simple,  $\mu_{i+1} \neq \mu_j$ . We observe that  $0 < a_i < a_{i+1}$ , i.e., in the mapping of  $x$  to  $w$  the coefficient in front of  $x_i$  changes by a smaller absolute value compared to the change in the coefficient in front of  $x_{i+1}$ . Thus,  $\mu(x) > \mu(a_i c_i x_i + a_{i+1} c_{i+1} x_{i+1}) \geq \mu(w)$  using the monotonicity of the Rayleigh quotient in the absolute values of the coefficients of the eigenvector expansion of its argument, which contradicts  $\mu(w) > \mu(x)$  proved in Theorem 2.2.  $\square$



has the vector-valued solution  $y(t)$ , which preserves the norm of the initial vector  $w$  since  $d\|y(t)\|^2/dt = 2(y(t), y'(t)) = 0$  as  $(y, \nabla\mu(y)) = 0$ . Without loss of generality we assume  $\|w\| = 1 = \|y(t)\|$ . The Rayleigh quotient function  $\mu(y(t))$  is decreasing since

$$\frac{d}{dt}\mu(y(t)) = (\nabla\mu(y(t)), y'(t)) = \left( \nabla\mu(y(t)), -\frac{\nabla\mu(y(t))}{\|\nabla\mu(y(t))\|} \right) = -\|\nabla\mu(y(t))\| \leq 0.$$

As  $\mu(y(0)) = \mu(w) < \mu_i$ , the function  $\mu(y(t))$  is strictly decreasing at least until it reaches  $\kappa > \mu_{i+1}$  as there are no eigenvalues in the interval  $[\kappa, \mu(y(0))] \subset (\mu_{i+1}, \mu_i)$ , but only eigenvectors can be special points of ODE (2.4). The condition  $\mu(y(\bar{t})) = \kappa$  thus uniquely determines  $\bar{t}$  for a given initial value  $w$ . The absolute value of the decrease of the Rayleigh quotient along the path  $L := \{y(t), 0 \leq t \leq \bar{t}\}$  is

$$\mu(w) - \kappa = \mu(y(0)) - \mu(y(\bar{t})) = \int_0^{\bar{t}} \|\nabla\mu(y(t))\| dt > 0.$$

Our continuation method (2.4) using the *normalized* gradient flow is nonstandard, but its advantage is that it gives the following simple expression for the length of  $L$ ,  $\text{Length}(L) = \int_0^{\bar{t}} \|y'(t)\| dt = \int_0^{\bar{t}} 1 dt = \bar{t}$ .

Since the initial value  $w$  is determined by  $x$ , we compare a generic  $x$  with the special choice  $x = x^* \in \text{span}\{x_i, x_{i+1}\}$ , using the superscript  $*$  to denote all quantities corresponding to the choice  $x = x^*$ . By Theorem 2.4  $x^* \in \text{span}\{x_i, x_{i+1}\}$  implies  $w^* \in \text{span}\{x_i, x_{i+1}\}$ , so we have  $y^*(t) \in \text{span}\{x_i, x_{i+1}\}$ ,  $0 \leq t \leq \bar{t}^*$  as  $\text{span}\{x_i, x_{i+1}\}$  is an invariant subspace for the gradient of the Rayleigh quotient. At the end points,  $\mu(y(\bar{t})) = \kappa = \mu(x) = \mu(y^*(\bar{t}^*))$ , by their definition. Our goal is to bound the initial value  $\mu(w^*) = \mu(y^*(0))$  by  $\mu(w) = \mu(y(0))$ , so we compare the lengths of the corresponding paths  $L^*$  and  $L$  and the norms of the gradients along these paths.

We start with the lengths. We obtain  $\phi_1(x^*) - \phi_\gamma(x^*) \leq \phi_1(x) - \phi_\gamma(x)$  by Theorem 2.5. Here the angle  $\phi_1(x) - \phi_\gamma(x)$  is the smallest angle between any two vectors on the cones boundaries  $\partial C_{\phi_\gamma(x)}(Bx)$  and  $\partial C_{\phi_1(x)}(Bx)$ . Thus,  $\phi_1(x) - \phi_\gamma(x) \leq \angle\{y(0), y(\bar{t})\}$  as our one vector  $y(0) = w \in \partial C_{\phi_\gamma(x)}(Bx)$  by Theorem 2.4, while the other vector  $y(\bar{t})$  cannot be inside the cone  $C_{\phi_\gamma(x)}(Bx)$  since  $\mu(w) > \kappa = \mu(y(\bar{t}))$  by Theorem 2.2. As  $y(t)$  is a unit vector,  $\angle\{y(0), y(\bar{t})\} \leq \text{Length}(L) = \bar{t}$  as the angle is the length of the arc—the shortest curve from  $y(0)$  to  $y(\bar{t})$  on the unit ball.

For our special  $*$ -choice, inequalities from the previous paragraph turn into equalities, as  $y^*(t)$  is in the intersection of the unit ball and the subspace  $\text{span}\{x_i, x_{i+1}\}$ , so the path  $L^*$  is the arc between  $y^*(0)$  to  $y^*(\bar{t}^*)$  itself. Combining everything together,

$$\begin{aligned} \bar{t}^* = \text{Length}(L^*) &= \angle\{y^*(0), y^*(\bar{t}^*)\} = \angle\{w^*, x^*\} = \varphi_1(x^*) - \varphi_\gamma(x^*) \\ &\leq \varphi_1(x) - \varphi_\gamma(x) \leq \angle\{y(0), y(\bar{t})\} \leq \text{Length}(L) = \bar{t}. \end{aligned}$$

By Theorem 2.5 on the norms of the gradient,  $-\|\nabla\mu(y^*(t^*))\| \geq -\|\nabla\mu(y(t))\|$  for each pair of independent variables  $t^*$  and  $t$  such that  $\mu(y^*(t^*)) = \mu(y(t))$ . Using Theorem 3.1, we conclude that  $\mu(w^*) = \mu(y^*(0)) \leq \mu(y(\bar{t} - \bar{t}^*)) \leq \mu(y(0)) = \mu(w)$  as  $\bar{t} - \bar{t}^* \geq 0$ , i.e., the subspace  $\text{span}\{x_i, x_{i+1}\}$  gives the smallest value  $\mu(w)$ .  $\square$

By Theorem 2.6 the poorest convergence is attained with  $x \in \text{span}\{x_i, x_{i+1}\}$  and with the corresponding minimizer  $w \in \text{span}\{x_i, x_{i+1}\}$  described in Theorem 2.4, so finally our analysis is now reduced to the two-dimensional space  $\text{span}\{x_i, x_{i+1}\}$ .

**THEOREM 2.7.** *Bound (2.3) holds and is sharp for  $x \in \text{span}\{x_i, x_{i+1}\}$ .*

*Proof.* Assuming  $\|x\| = 1$  and  $\|x_i\| = \|x_{i+1}\| = 1$ , we derive

$$(2.5) \quad |(x, x_i)|^2 = \frac{\mu(x) - \mu_{i+1}}{\mu_i - \mu_{i+1}} > 0 \text{ and } |(x, x_{i+1})|^2 = \frac{\mu_i - \mu(x)}{\mu_i - \mu_{i+1}},$$

and similarly for  $w \in \text{span}\{x_i, x_{i+1}\}$  where  $(B + \alpha I)w = Bx$ .

Since  $B > 0$ , we have  $x = (I + \alpha B^{-1})w$ . Assuming  $\alpha = -\mu_{i+1}$ , this identity implies  $x = x_i$ , which contradicts our assumption that  $x$  is not an eigenvector. For  $\alpha \neq -\mu_{i+1}$  and  $\alpha > -\mu_i$  by Theorem 2.4, the inverse  $(B + \alpha I)^{-1}$  exists.

Next we prove that  $\alpha > 0$  and that it is a strictly decreasing function of  $\kappa := \mu(x) \in (\mu_{i+1}, \mu_i)$ . Indeed, using  $Bx = (B + \alpha I)w$  and our cosine-based definition of the angles, we have  $0 < (w, (B + \alpha I)w)^2 = (w, Bx)^2 = \|w\|^2 \|Bx\|^2 \cos^2 \phi_\gamma(x)$ , where  $\|Bx\|^2 \cos^2 \phi_\gamma(x) = \|Bx\|^2 - \gamma^2 \|Bx - \kappa x\|^2$ . We substitute  $w = (B + \alpha I)^{-1}Bx$ , which gives  $((B + \alpha I)^{-1}Bx, Bx)^2 = \|(B + \alpha I)^{-1}Bx\|^2 (\|Bx\|^2 - \gamma^2 \|Bx - \kappa x\|^2)$ . Using (2.5), multiplication by  $(\mu_i + \alpha)^2(\mu_{i+1} + \alpha)^2$  leads to a simple quadratic equation,  $a\alpha^2 + b\alpha + c = 0$ ,  $a = \gamma^2(\kappa(\mu_i + \mu_{i+1}) - \mu_i\mu_{i+1})$ ,  $b = 2\gamma^2\kappa\mu_i\mu_{i+1}$ ,  $c = -(1 - \gamma^2)\mu_i^2\mu_{i+1}^2$  for  $\alpha$ . As  $a > 0$ ,  $b > 0$ , and  $c < 0$ , the discriminant is positive and the two solutions for  $\alpha$ , corresponding to the minimum and maximum of the Rayleigh quotient on  $C_{\phi_\gamma(x)}(Bx)$ , have different signs. The proof of Theorem 2.4 analyses the direction of the gradient of the Rayleigh quotient to conclude that  $\beta > 0$  and  $\alpha > -\mu(w)$  correspond to the minimum. Repeating the same arguments with  $\beta < 0$  shows that  $\alpha < -\mu(w)$  corresponds to the maximum. But  $\mu(w) > 0$  since  $B > 0$ , hence the negative  $\alpha$  corresponds to the maximum and thus the positive  $\alpha$  corresponds to the minimum. We observe that the coefficients  $a > 0$  and  $b > 0$  are evidently increasing functions of  $\kappa \in (\mu_{i+1}, \mu_i)$ , while  $c < 0$  does not depend on  $\kappa$ . Thus  $\alpha > 0$  is strictly decreasing in  $\kappa$ , and taking  $\kappa \rightarrow \mu_i$  gives the smallest  $\alpha = \mu_{i+1}(1 - \gamma)/\gamma > 0$ .

Since  $(B + \alpha I)w = Bx$  where now  $\alpha > 0$ , condition  $(x, x_i) \neq 0$  implies  $(w, x_i) \neq 0$  and  $(x, x_{i+1}) = 0$  implies  $(w, x_{i+1}) = 0$ , so we introduce the convergence factor as

$$\sigma^2(\alpha) := \frac{\mu_i - \mu(w)}{\mu(w) - \mu_{i+1}} \frac{\mu(x) - \mu_{i+1}}{\mu_i - \mu(x)} = \left| \frac{(w, x_{i+1})}{(w, x_i)} \right|^2 \left| \frac{(x, x_i)}{(x, x_{i+1})} \right|^2 = \left( \frac{\mu_{i+1} - \mu_i + \alpha}{\mu_i - \mu_{i+1} + \alpha} \right)^2,$$

where we use (2.5) and again  $(B + \alpha I)w = Bx$ . We notice that  $\sigma(\alpha)$  is a strictly decreasing function of  $\alpha > 0$  and thus takes its largest value for  $\alpha = \mu_{i+1}(1 - \gamma)/\gamma$  giving  $\sigma = \gamma + (1 - \gamma)\mu_{i+1}/\mu_i$ , i.e., bound (2.3) that we are seeking.

The convergence factor  $\sigma^2(\alpha)$  cannot be improved without introducing extra terms or assumptions. But  $\sigma^2(\alpha)$  deals with  $w \in C_{\phi_\gamma(x)}(Bx)$ , not with the actual iterate  $x'$ . We now show that for  $\kappa \in (\mu_{i+1}, \mu_i)$  there exist a vector  $x \in \text{span}\{x_i, x_{i+1}\}$  and a preconditioner  $T$  satisfying (2.2) such that  $\kappa = \mu(x)$  and  $x' \in \text{span}\{w\}$  in both real and complex cases. In the complex case, let us choose  $x$  such that  $\mu(x) = \kappa$  and  $x = |x|$  according to (2.5), then the real vector  $w = |w| \in C_{\phi_\gamma(x)}(Bx)$  is a minimizer of the Rayleigh quotient on  $C_{\phi_\gamma(x)}(Bx)$ , since  $\mu(w) = \mu(|w|)$  and  $|(w, B|x|)| \leq (|w|, B|x|)$ .

Finally, for a real  $x$  with  $\mu(x) = \kappa$  and a real properly scaled  $y \in C_{\phi_\gamma(x)}(Bx)$  there is a real matrix  $T$  satisfying (2.2) such that  $y = Bx - (I - T)(Bx - \kappa x)$ , which leads to (2.1) with  $\mu(x)x' = y$ . Indeed, for the chosen  $x$  we scale  $y \in C_{\phi_\gamma(x)}(Bx)$  such that  $(y, Bx - y) = 0$  so  $\|Bx - y\| = \sin \phi_\gamma(x) \|Bx\| = \gamma \|Bx - \kappa x\|$ . As vectors  $Bx - y$  and  $\gamma(Bx - \kappa x)$  are real and have the same length there exists a *real* Householder reflection  $H$  such that  $Bx - y = H\gamma(Bx - \kappa x)$ . Setting  $T = I - \gamma H$  we obtain the required identity. Any Householder reflection is symmetric and has only two distinct eigenvalues  $\pm 1$ , so we conclude that  $T$  is real symmetric (and thus Hermitian in the complex case) and satisfies (2.2).  $\square$

**3. Appendix.** The integration of inverse functions theorem follows.

**THEOREM 3.1.** *Let  $f, g : [0, b] \rightarrow \mathbf{R}$  for  $b > 0$  be strictly monotone increasing smooth functions and suppose that for  $a \in [0, b]$  we have  $f(a) = g(b)$ . If for all*

$\alpha, \beta \in [0, b]$  with  $f(\alpha) = g(\beta)$  the derivatives satisfy  $f'(\alpha) \leq g'(\beta)$ , then for any  $\xi \in [0, a]$  we have  $f(a - \xi) \geq g(b - \xi)$ .

*Proof.* For any  $\xi \in [0, a]$  we have (using  $f(a) = g(b)$ )

$$\xi = \int_{g(b-\xi)}^{g(b)} (g^{-1})'(y) dy = \int_{f(a-\xi)}^{g(b)} (f^{-1})'(y) dy.$$

If  $y = f(\alpha) = g(\beta)$ , then for the derivatives of the inverse functions it holds that  $(g^{-1})'(y) \leq (f^{-1})'(y)$ . Since  $f$  and  $g$  are strictly monotone increasing functions the integrands are positive functions and  $g(b - \xi) < g(b)$  as well as  $f(a - \xi) < f(a) = g(b)$ . Comparing the lower limits of the integrals gives the statement of the theorem.  $\square$

**Conclusions.** We present a new geometric approach to the convergence analysis of a preconditioned fixed-step gradient eigensolver which reduces the derivation of the convergence rate bound to a two-dimensional case. The main novelty is in the use of a continuation method for the gradient flow of the Rayleigh quotient to locate the two-dimensional subspace corresponding to the smallest change in the Rayleigh quotient and thus to the slowest convergence of the gradient eigensolver.

An elegant and important result such as Theorem 1.1 should ideally have a textbook-level proof. We have been trying, unsuccessfully, to find such a proof for several years, so its existence remains an open problem.

**Acknowledgments.** We thank M. Zhou of University of Rostock, Germany for proofreading. M. Argentati of University of Colorado Denver, E. Ovtchinnikov of University of Westminster, and anonymous referees have made numerous great suggestions to improve the paper and for future work.

#### REFERENCES

- [1] J. H. BRAMBLE, J. E. PASCIAK, AND A. V. KNYAZEV, *A subspace preconditioning algorithm for eigenvector/eigenvalue computation*, Adv. Comput. Math., 6 (1996), pp. 159–189.
- [2] E. G. D'YAKONOV, *Optimization in solving elliptic problems*, CRC Press, 1996.
- [3] A. V. KNYAZEV, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541.
- [4] A. V. KNYAZEV, *Preconditioned eigensolvers: practical algorithms*, In Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, editors, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, pp. 352–368. SIAM, Philadelphia, 2000.
- [5] A. V. KNYAZEV, *Computation of eigenvalues and eigenvectors for mesh problems: algorithms and error estimates*, (In Russian), Dept. Num. Math., USSR Ac. Sci., Moscow, 1986.
- [6] A. V. KNYAZEV, *Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem*, Russian J. Numer. Anal. Math. Modelling, 2 (1987), pp. 371–396.
- [7] A. V. KNYAZEV, *Preconditioned eigensolvers—an oxymoron?*, Electron. Trans. Numer. Anal., 7 (1998), pp. 104–123.
- [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 (2003), pp. 95–114.
- [9] K. NEYMEYR, *A geometric theory for preconditioned inverse iteration. I: Extrema of the Rayleigh quotient*, Linear Algebra Appl., 322 (2001), pp. 61–85.
- [10] K. NEYMEYR, *A geometric theory for preconditioned inverse iteration. II: Convergence estimates*, Linear Algebra Appl., 322 (2001), pp. 87–104.