A Hierarchy of
Preconditioned Eigensolvers
for Elliptic Differential Operators

Habilitationsschrift
von
Dr. Klaus Neymeyr

Mathematische Fakultät
Universität Tübingen

September 2001

V02/02WWW

CONTENTS

1 Introduction 1

1.1 Mesh eigenproblems for elliptic differential operators 4

1.1.1 A model problem and its discretization 4

1.1.2 Gradient type eigensolvers 8

1.1.3 Preconditioned eigensolvers 9

1.1.4 Justification of preconditioning for eigensolvers 12

1.1.5 Multigrid and multilevel preconditioning 16

1.1.6 Multigrid solvers for elliptic eigenproblems 18

1.2 A hierarchy of preconditioned eigensolvers 20

1.2.1 Subspace iteration 21

1.2.2 Preconditioned subspace iteration 23

1.2.3 Subspace iteration and gradient type schemes 25

1.2.4 Generalized inexact solvers 26

1.3 An application: Electronic structure theory 28

1.4 Model analysis of inverse iteration 30

2 Preconditioned inverse iteration 37

2.1 Preconditioning for eigenvalue solvers 37

2.1.1 Assumptions on the preconditioners 39

2.1.2 Normal form of preconditioners 40

2.2 A geometric representation 43

2.3 Multiple eigenvalues 45

2.4 A convergence theorem 47

3 Analysis of fastest convergence 53

3.1 Best preconditioning 54

3.1.1 Extremum points in $E_\gamma(c)$ 55

3.1.2 A necessary condition for infimum points 58

3.1.3 Parametrization of infimum points 60

3.1.4 Dependence of the shift parameter α on γ 61

3.1.5 Bifurcation of the infima curve 63

3.2 Extremal quantities on $L(\lambda)$ 64

3.2.1 Extrema of $\|\nabla\lambda(c)\|$ 65

3.2.2 Extremal properties of the cone $C_\gamma(c)$ 66

3.2.3 Angle dependence of the Rayleigh quotient on $C_\gamma(c)$ 67

3.3 Mini-dimensional convergence analysis . . . 69
3.4 Convergence estimates for the Rayleigh quotient . . . 72
3.5 Critical conclusion . . . 77

4 Concise convergence estimates 79
4.1 Reformulation of the mini-dimensional analysis . . . 80
4.2 A matrix function approach . . . 82
4.2.1 An abstract convergence estimate . . . 82
4.2.2 Several estimates on extremal convergence . . . 84
4.3 A critical comparison of convergence estimates . . . 89

5 A preconditioned subspace eigensolver 93
5.1 Analysis of simplified subspace solvers . . . 93
5.2 A convergence theorem . . . 96
5.3 Generalizations and remarks . . . 97

6 PINVIT(2) – Preconditioned steepest descent 99
6.1 Rayleigh-Ritz accelerates convergence . . . 100
6.1.1 Steepest descent for the eigenproblem . . . 102
6.2 Convergence analysis of INVIT(2) . . . 103
6.3 PINVIT(2) convergence theory . . . 112
6.3.1 Elementary results on extremal convergence . . . 112
6.3.2 The c-basis representation . . . 114
6.3.3 A geometric representation . . . 115
6.3.4 A conjecture on the subspace of poorest convergence . . . 117
6.4 Mini-dimensional analysis . . . 118
6.5 Numerical algorithms . . . 123
6.5.1 Numerical results for the Laplacian . . . 123
6.5.2 Connectedness of $L_+(\lambda)$ . . . 127
6.5.3 A search algorithm on $L_+(\lambda)$ . . . 128
6.5.4 On conjecture 6.21 . . . 131
6.6 Critical conclusion . . . 131

7 Numerical experiments 133
7.1 Comparison of the PINVIT(k,s) schemes . . . 133
7.2 An adaptive subspace eigensolver . . . 140

8 Conclusion and outlook 143

References 147

List of Symbols 157

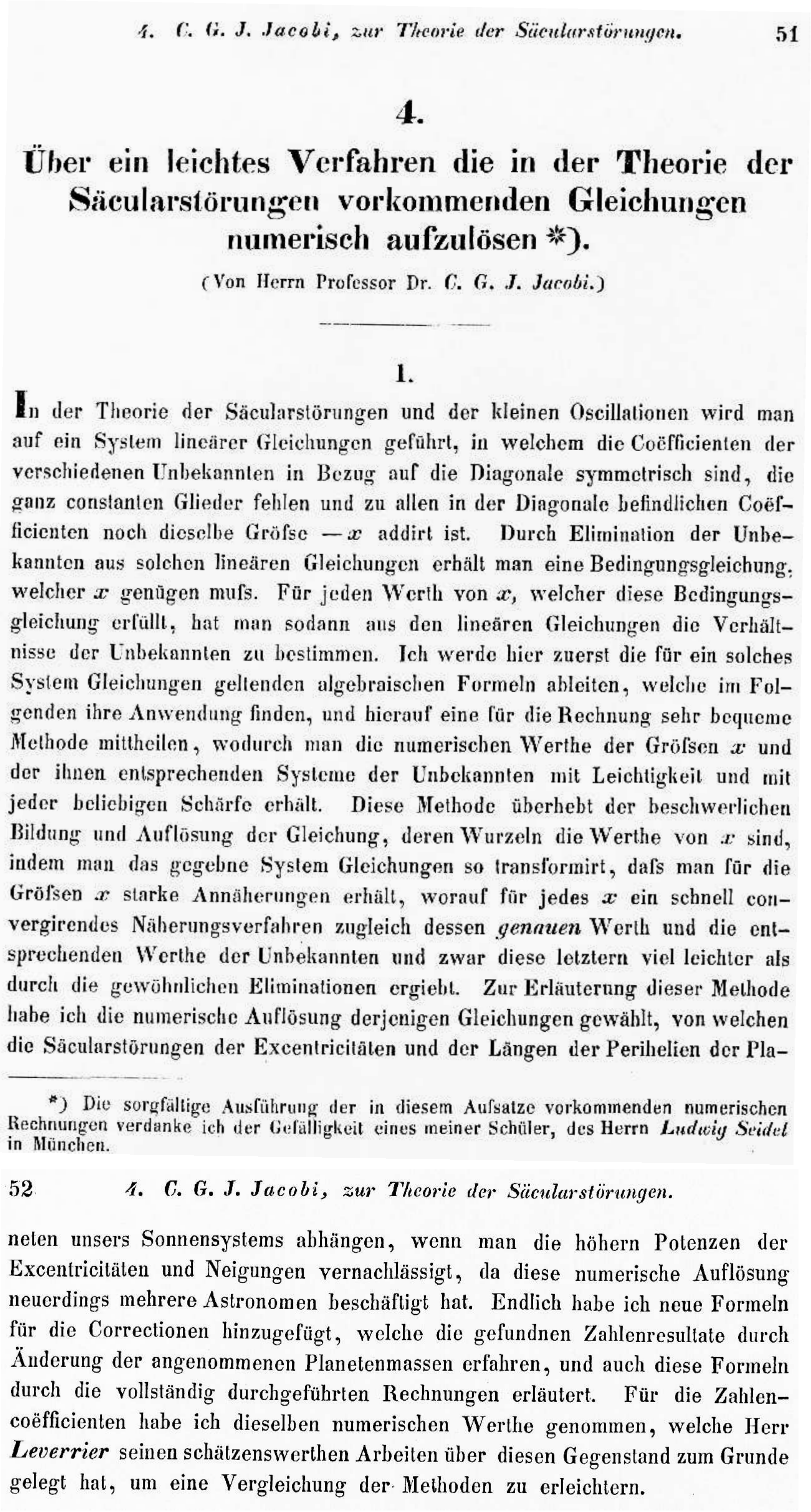
1. Introduction

C.G.J. Jacobi's work from 1846 in [64] entitled "Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen", marks the beginning of the research on the numerical solution of eigenproblems, which even after 155 years remains an important and vital area in numerical linear algebra. Jacobi investigated a small eigenvalue problem within only 7 variables describing the stability of the orbits of the 7 planets which were known at that time. The necessary numerical computations were done by L. Seidel. A facsimile of the introduction to Jacobi's paper is shown in Figure 1.1.

The next milestone in the numerical solution of eigenvalue problems were the five papers written by Wielandt in 1943 and 1944 [135–139], where he introduced the power method and inverse iteration for computing eigenfunctions of linear operators. The expenditure of work for such eigenvalue/vector computations was high. In order to determine the first eigenvalue (of largest modulus) of a complex 4×4 matrix, Wielandt needed about 50–80 minutes using an electro-mechanical calculator with a six digit accuracy.

Since the computation of eigenvalues and eigenvectors is significantly more complicated than the solution of linear systems, it is not surprising that the number of those early works on numerical algorithms is limited. Before 1940, computing eigenvalues was most often based on computing the characteristic polynomial and finding its roots. In contrast to this, the spectral theory for partial differential operators was in those days in a much more advanced state, as reflected for instance by the monograph of Courant and Hilbert [26]; see also the book of Collatz [25], which additionally contains numerous examples of eigenvalue problems for mechanical systems.

Around the year 1950, with the advent of electronic computers, the situation changed drastically as substantial progress was made in the numerical solution of the eigenvalue problem and its error analysis. The pioneering work was done by researchers like Arnoldi, Bauer, Francis, Givens, Householder, Kublanovskaya, Lanczos, Rutishauser, Wilkinson and several others. The state of the art at the beginning of the 1960s is summarized in the monograph of Wilkinson [141], which still constitutes an important reference. Many of the algorithms in this early phase are based on matrix transforming techniques like the very successful QR algorithm, due to Francis and Kublanovskaya. The QR algorithm is most frequently used for the calculation of the set of eigenvalues of general but relatively small matrices.



4. C. G. J. Jacobi, zur Theorie der Säcularstörungen. 51

4.

Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen *).

(Von Herrn Professor Dr. *C. G. J. Jacobi.*)

1.

In der Theorie der Säcularstörungen und der kleinen Oscillationen wird man auf ein System linearer Gleichungen geführt, in welchem die Coëfficienten der verschiedenen Unbekannten in Bezug auf die Diagonale symmetrisch sind, die ganz constanten Glieder fehlen und zu allen in der Diagonale befindlichen Coëfficienten noch dieselbe Gröſse $-x$ addirt ist. Durch Elimination der Unbekannten aus solchen linearen Gleichungen erhält man eine Bedingungsgleichung, welcher x genügen muſs. Für jeden Werth von x, welcher diese Bedingungsgleichung erfüllt, hat man sodann aus den linearen Gleichungen die Verhältnisse der Unbekannten zu bestimmen. Ich werde hier zuerst die für ein solches System Gleichungen geltenden algebraischen Formeln ableiten, welche im Folgenden ihre Anwendung finden, und hierauf eine für die Rechnung sehr bequeme Methode mittheilen, wodurch man die numerischen Werthe der Gröſsen x und der ihnen entsprechenden Systeme der Unbekannten mit Leichtigkeit und mit jeder beliebigen Schärfe erhält. Diese Methode überhebt der beschwerlichen Bildung und Auflösung der Gleichung, deren Wurzeln die Werthe von x sind, indem man das gegebne System Gleichungen so transformirt, daſs man für die Gröſsen x starke Annäherungen erhält, worauf für jedes x ein schnell convergirendes Näherungsverfahren zugleich dessen *genauen* Werth und die entsprechenden Werthe der Unbekannten und zwar diese letztern viel leichter als durch die gewöhnlichen Eliminationen ergiebt. Zur Erläuterung dieser Methode habe ich die numerische Auflösung derjenigen Gleichungen gewählt, von welchen die Säcularstörungen der Excentricitäten und der Längen der Perihelien der Planeten unsers Sonnensystems abhängen, wenn man die höhern Potenzen der Excentricitäten und Neigungen vernachlässigt, da diese numerische Auflösung neuerdings mehrere Astronomen beschäftigt hat. Endlich habe ich neue Formeln für die Correctionen hinzugefügt, welche die gefundnen Zahlenresultate durch Änderung der angenommenen Planetenmassen erfahren, und auch diese Formeln durch die vollständig durchgeführten Rechnungen erläutert. Für die Zahlencoëfficienten habe ich dieselben numerischen Werthe genommen, welche Herr *Leverrier* seinen schätzenswerthen Arbeiten über diesen Gegenstand zum Grunde gelegt hat, um eine Vergleichung der Methoden zu erleichtern.

*) Die sorgfältige Ausführung der in diesem Aufsatze vorkommenden numerischen Rechnungen verdanke ich der Gefälligkeit eines meiner Schüler, des Herrn *Ludwig Seidel* in München.

52 4. C. G. J. Jacobi, zur Theorie der Säcularstörungen.

Figure 1.1: *Introduction of Jacobi [64].*

Today, as mathematical models invade more and more disciplines, a plurality of problems in science and engineering leads to eigenproblems. At the same time, the dimension of the eigenproblems becomes larger and larger, which constitutes the necessity for more efficient eigensolvers which are capable of solving, e.g., those extremely large eigenvalue problems arising from the discretization of eigenvalue problems for partial differential operators. Of course, since the 1960s considerable progress has been made in the linear algebra of eigensolvers as described in the standard monographs by Parlett [107], Golub and van Loan [48], Saad [117] and the most recent "Templates for the Solution of Algebraic Eigenvalue Problems" [5]. For some survey of the history and main research developments in the area of computational methods of eigenvalue problems see Wilkinson [140] as well as Golub and van der Vorst [47, 129].

Nevertheless, there is still an important *challenge for modern eigensolvers*, namely to treat those *extremely large and sparse (generalized) eigenvalue problems*, which derive from the *discretization of partial differential operators*. Such solvers should feature *optimal complexity* even on non-uniform grids and *should not require any regularity assumptions*.

In the present work we take up this problem and analyze iterative solvers for mesh discretizations of eigenvalue problems for self-adjoint and coercive elliptic differential operators. The discretization can be done by means of the finite element or the finite difference method. The main difficulty with these discretized eigenproblems is their sheer size. Today, even on a standard personal computer, one wants to solve such problems for a number of variables up to several millions. Later, we will discuss in more detail why classical solvers like QR or Lanczos cannot be applied to these problems. Instead, special solvers are required which are capable of exploiting the structure of these problems stemming from discretized partial differential operators.

Another characteristic trait of these eigenproblems is that one is only interested in a small part of the spectrum, typically in the smallest eigenvalues or those nearest to some prescribed value. Often those eigenvalues have a practical physical meaning, e.g. they characterize the base frequencies of some vibrating mechanical structure modeled by an eigenvalue problem for an elliptic partial differential operator. The number of eigenvalues that are to be determined (together with the corresponding eigenfunctions), ranges from 1 up to several hundred.

The aim of this work is to develop a *new theoretical framework* for the efficient solution of such extremely large eigenproblems. A central element of the solvers presented here are *approximate inverses* of that operator whose eigenproblem is to be solved. One usually calls such an approximate inverse a *preconditioner*. In our setup very efficient preconditioners are available, which are based on multigrid iterations or on domain decomposition techniques. A major advantage of the present approach is that we are able to completely separate the questions of the construction of such an approximate inverse and that of the analysis of the iterative eigensolver. In this way we can treat the preconditioner as a "black box"; only specific constants describing the quality of the preconditioner enter into the analysis (in Section 1.1.3 we provide an analytic description of such quality conditions describing how well the preconditioner approximates the exact inverse of the given operator). Therefore, the proof techniques presented in the following are predominantly based on some type of geometric analysis of the eigensolver and its corresponding linear algebra.We derive several new sharp and non-asymptotic convergence estimates for such preconditioned eigensolvers. As a by-product, a new geometric interpretation of these schemes is suggested, providing the basis for a renewed understanding of the convergence analysis.

The resulting eigensolvers are not only conceptionally simple, easy to implement and cheap, but also, as an outcome of the convergence theory, robust and stable. Grid-independent convergence can be guaranteed.

This work is organized as follows: In the remaining part of this chapter we introduce and justify preconditioning for eigenproblems. Moreover, we suggest a unifying framework for preconditioned eigensolvers, in which they are derived systematically from some preconditioned variant of subspace iteration. Within this framework new sharp non-asymptotic convergence estimates for the most basic preconditioned eigensolver can be derived, see Chapter 2. Our new approach to the convergence analysis of preconditioned eigensolvers also allows to derive several estimates on the fastest possible convergence, as done in Chapter 3. These new sharp estimates give an explanation for the extremely fast convergence often observed in the first steps of the iteration. The mentioned estimates are somewhat complex and awkward; therefore we derive drastically simplified estimates in Chapter 4 without too much loss of sharpness. Chapter 5 treats the convergence theory of a basic preconditioned subspace eigensolver. In Chapter 6 we progress one step toward more advanced schemes (within the hierarchy of preconditioned eigensolvers) and provide a (partial) convergence analysis. Finally, in Chapter 7 we report on the results of some numerical experiments.

1.1 Mesh eigenproblems for elliptic differential operators

1.1.1 A model problem and its discretization

In order to introduce the above mentioned large and sparse generalized matrix eigenvalue problem, let us now consider the eigenvalue problem for a self-adjoint and coercive elliptic differential operator $\mathcal{L}$. We restrict the discussion to a second-order *model problem* in $\mathbb{R}^2$ or $\mathbb{R}^3$.

Therefore, let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded, open, connected set with a Lipschitz continuous boundary Γ, which has been subdivided in two disjoint sets Γ_1 and Γ_2. The problem is to find (some of the smallest) eigenvalues λ together with real-valued eigenfunctions $u = u(x)$ satisfying

$$\begin{aligned} -\nabla(c(x)\nabla u) + q(x)u &= \lambda u, \qquad && x \in \Omega, \\ u &= 0, && x \in \Gamma_1, \\ \nu \cdot c(x)\nabla u &= 0, && x \in \Gamma_2. \end{aligned} \tag{1.1}$$

<table>
<tr><td colspan="2">Let $\mathcal{L}$ be a self-adjoint and coercive elliptic partial differential operator.</td></tr>
<tr><td colspan="2">WEAK FORMULATION OF THE</td></tr>
<tr><td>BOUNDARY VALUE PROBLEM:</td><td>EIGENVALUE PROBLEM:</td></tr>
<tr><td>Find $u \in \mathcal{H}$ with</td><td>Find $(u, \lambda) \in \mathcal{H} \times \mathbb{R}$</td></tr>
<tr><td>$a(u, v) = (f, v), \quad v \in \mathcal{H},$</td><td>$a(u, v) = \lambda(u, v), \quad v \in \mathcal{H},$</td></tr>
<tr><td colspan="2">for the bilinear form $a(\cdot, \cdot)$ associated with $\mathcal{L}$, the inner product $(\cdot, \cdot)$ and an appropriate Hilbert space $\mathcal{H}$.</td></tr>
<tr><td colspan="2">MESH DISCRETIZATION OF THE VARIATIONAL PROBLEM</td></tr>
<tr><td>Linear system:</td><td>Generalized matrix eigenproblem:</td></tr>
<tr><td>$Ax = b, \quad A \in \mathbb{R}^{n \times n},$</td><td>$Ax = \lambda Mx, \quad A, M \in \mathbb{R}^{n \times n},$</td></tr>
<tr><td>$A > 0, \quad A^T = A.$</td><td>$A > 0, \quad A^T = A,$</td></tr>
<tr><td></td><td>$M \geq 0, \quad M^T = M.$</td></tr>
<tr><td colspan="2">EFFICIENT SOLVERS</td></tr>
<tr><td>• Multigrid solvers,</td><td>• Direct multigrid eigensolvers,</td></tr>
<tr><td>• Multigrid preconditioned conjugate gradient schemes.</td><td>• <b>Multigrid preconditioned eigensolvers; PINVIT(k,s) schemes.</b></td></tr>
<tr><td colspan="2">TOTAL COMPLEXITY OF THE ITERATIVE SOLVER</td></tr>
<tr><td>$\mathcal{O}(n)$ to compute x.</td><td>$\mathcal{O}(n)$ to determine (λ_1, x_1).</td></tr>
<tr><td colspan="2">Grid independent convergence for high-quality multigrid preconditioners;
No regularity assumptions required for the best multilevel preconditioners.</td></tr>
</table>

Table 1.1: Solution of boundary value and eigenvalue problems for elliptic operators.

Therein, ν denotes the exterior unit normal to Γ_2 and $c(x)$ is a symmetric positive definite matrix-valued function while $q(x)$ is assumed to be a real-valued and, for the sake of having only positive eigenvalues, a positive function. Both functions are assumed piecewise continuous.

As a next step toward a solution of (1.1), one can derive its weak formulation; Table 1.1 contains a schematic description. The variational form allows the application of the mathematically sound spectral theory for self-adjoint compact operators in Hilbert spaces, which guarantees the existence of a countable set of eigenvalues—each eigenvalue corresponding to a finite dimensional invariant subspace. We do not go into the details but refer to Raviart and Thomas [112] as well as Babuška and Osborn [3].

The finite-dimensional Galerkin approximation of the weak form (or its Rayleigh-Ritz discretization) leads to the generalized matrix eigenvalue problem

$$Ax = \lambda Mx, \tag{1.2}$$

where $A \in \mathbb{R}^{n\times n}$ and $M \in \mathbb{R}^{n\times n}$ are symmetric positive (semi)definite matrices, which are very large and sparse. Then A is called the discretization (or stiffness) matrix and M the mass matrix.

To introduce some necessary notation let (λ_i, x_i) be the eigenpairs of (A, M), which are assumed to satisfy

$$(x_i, Ax_j) = \lambda_i\delta_{ij}, \qquad (x_i, Mx_j) = \delta_{ij},$$

where δ_{ij} denotes the Kronecker delta. The n real positive eigenvalues λ_i of (A, M) may have arbitrary multiplicity and are in such an order that

$$0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n.$$

The dimension n of the basis of finite element functions may exceed 10^7 even on standard, present-day personal computers. Obviously, the sparse matrices A and M cannot be stored in the computer as full matrices. As typical of finite element codes, they are either stored in some sparse matrix format, or program routines are provided to compute the matrix-vector products Ax and Mx by local compilation for any input vector x. Note that both matrices have only a small number of nonzero elements per row, e.g., typically 5 or 9 elements for a linear finite element discretization of a second order elliptic differential operator in $\mathbb{R}^2$ or $\mathbb{R}^3$.

For our needs, there is no necessity to force M to equal a diagonal matrix—as sometimes done in the finite element method and called mass-lumping. Surely, most of the available eigensolvers are designed to solve the standard matrix eigenvalue problem and some linear transformation is required to reduce (1.2) to the standard form. But the preconditioned eigensolvers, as presented in this work, can be applied to the standard and to the generalized eigenvalue problem with only marginal changes. Nevertheless, for the sake of convenience, we restrict the analysis in Chapters 2 to 6 to the standard eigenvalue problem. The theoretical justification of such a reduction is a change of the inner product, see [73, 97] for the analysis.

There is a further important feature of the matrix pencil (A, M) to be mentioned: The condition number of A is typically very large. For second-order partial differential operators (like problem (1.1)) it increases like h^{-2} in the mesh-width h. In contrast to this, the condition number of M is uniformly bounded by a constant independent of the mesh size.

Under the given restrictions the set of efficient solvers for (1.2) is small. Most of the "linear algebra textbook eigensolvers" cannot be applied for the following reasons:

1. Lack of computer storage:
   The large, sparse matrices A and M can neither be *factored* (like the LU or Cholesky decomposition), nor be iteratively *transformed* (e.g. by successive similarity transformations), since the computer storage for holding the typically denser computed factors or transformed matrices is not available. Usually the whole available storage is needed to hold the high-dimensional iteration vector (or several of them in the case of a subspace iteration) and in order to define the (sometimes adaptively generated non-uniform) grid structure. Within a matrix-free environment the generated mesh defines implicitly how to evaluate the products Ax and Mx, which is done within specific subroutines. The given restrictions exclude, among others, not only inverse iteration and the Rayleigh quotient iteration with their variants, but also the QR-method and the Jacobi iteration.

2. Ill-conditioning of A:
   The large condition number of A, e.g. for the discrete Laplacian $\kappa(A) \simeq h^{-2}$, impedes the successful application of the Lanczos process, as the convergence slows down considerably for decreasing h. See [90] for a combination of preconditioning and the Lanczos algorithm. Nevertheless, for problems of moderate size the popular package ARPACK [79] can provide reliable numerical results.

   A large condition number can also destabilize the solution of linear systems in A (but a direct solution has already been excluded by the points made above).

The question now is how to overcome these difficulties in order to construct efficient eigensolvers? Preconditioning (in other words, the application of an approximate inverse) can provide some cure.

To this end let us first derive the basic iterative scheme of gradient and preconditioned eigensolvers in Sections 1.1.2 and 1.1.3. A brief survey on various approaches to preconditioning for eigensolvers is contained in Section 1.1.4. Subsequently, we review the elements of multigrid preconditioning in Section 1.1.5; the deep results developed in the early 1990s in the field of *multilevel preconditioning techniques* provide the theoretical basis for making possible preconditioned eigensolvers with *optimal computational complexity* on *non-uniform* grids and *without any requirements on the regularity*. Finally, in Section 1.1.6 an outline of alternative multigrid schemes for the eigenproblem is given.

1.1.2 Gradient type eigensolvers

Our goal is to develop *fast* (in the best case with a total complexity of $\mathcal{O}(n)$) and *storage-efficient* iterative solvers for the *partial eigenproblem* (1.2). A partial eigenproblem means that we are only interested in parts of the spectrum. Here we try to compute only a modest number of the smallest eigenvalues together with the corresponding invariant subspace. To sum up, the eigenproblem (1.2) is given for the (sometimes extremely) large and sparse matrices A and M. Moreover, A is an ill-conditioned matrix. Under the given restrictions it is highly undesirable to factor these matrices or any linear combination of them into a product of matrices. In other words, we avoid any direct solution of equations involving these matrices because of its computational costs and the usually unavailable storage.

On the other hand we should compile the minimal set of operations which we are willing to make available for an eigensolver. For arbitrary $x, y \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ these are:

MINIMAL SET OF AVAILABLE OPERATIONS:

1. Matrix-vector multiplications: $x \to Ax$ and $x \to Mx$,
2. Linear operations: $x + y$, λx,
3. Inner products: $(x, y) = x^T y$.

The matrix-vector multiplications can be realized by $\mathcal{O}(n)$ floating point operations since the number of nonzero elements per row of A and M is a small fixed number. Therefore, any fixed combination of the listed operations can be done for total costs that behave like $\mathcal{O}(n)$.

Even with this small number of operations one can construct a first, preliminary eigensolver, which allows to determine the smallest eigenvalue of A together with an eigenvector: The idea is to reformulate the eigenvalue problem (1.2) as an *optimization problem* for the (generalized) Rayleigh quotient

$$\lambda := \lambda(x) = \frac{(x, Ax)}{(x, Mx)}. \tag{1.3}$$

As suggested by Kantorovich [65] as well as Hestenes and Karush [58] one tries to correct a given iterate x in the direction of the negative gradient of the Rayleigh quotient in order to decrease the Rayleigh quotient of the new iterate.

As the gradient of the Rayleigh quotient (1.3) is given by

$$\nabla\lambda(x) = \frac{2}{(x, Mx)}(Ax - \lambda(x)Mx), \tag{1.4}$$

the so-called *gradient method* for the eigenproblem has the form

$$x' := x - \omega(Ax - \lambda(x)Mx). \tag{1.5}$$

Therein, ω is a scaling parameter, which has to be determined appropriately. In the best (but most expensive) case, an optimal parameter $\hat{\omega}$ is determined in a way that the Rayleigh quotient of the new iterate x' is minimized, i.e.

$$\hat{\omega} \in \arg\min_{\omega\in\mathbb{R}} \lambda(x - \omega(Ax - \lambda Mx)).$$

This scheme is called *steepest descent* for the eigenvalue problem. The computation of $\hat{\omega}$ can be done implicitly by applying the Rayleigh-Ritz method to the 2D subspace spanned by x and $r := Ax - \lambda(x)Mx$. As long as the residual r is a nonzero vector, i.e. that x is no eigenvector of A, the Rayleigh quotients of the iterates of the gradient method form a sequence of decreasing numbers, which (usually) tends to the smallest eigenvalue λ_1. Then the iterates themselves converge to an associated eigenvector.

Unfortunately, the simple gradient method (1.5) and even steepest descent suffer from *poor convergence in the case of the ill-conditioned eigenproblem* under consideration. Hence, these schemes cannot satisfy the demand for grid-independent convergence. A direct proof of this fact is given in Section 3 of the Technical Report [94]. This disappointing result does not come as a surprise since the gradient method (1.5) and the Lanczos scheme span the same Krylov space. While the first scheme only extracts a suboptimal approximation from this Krylov space, even the latter scheme is known to suffer from ill-conditioning of A, cf. Section 1.1.1.

The gradient method and some acceleration techniques (e.g. by various scaling strategies or by adoption of the conjugate gradients method) are treated, for instance, in Bradbury and Fletcher [11], Faddeev and Faddeeva [39], McCormick [85, 86], Ruhe and Wiberg [115], Rodrigue [113], Longsine and McCormick [83], Döhler [30] as well as in Feng and Owen [42]. The very similar scheme of *steepest ascent*, in which the Rayleigh quotient of x' is maximized with respect to the choice of ω, has been analyzed by Knyazev and Skorokhodov [76], where also sharp convergence estimates have been derived.

1.1.3 Preconditioned eigensolvers

Preconditioning can improve the convergence properties of gradient type eigensolvers decisively. A *preconditioner* $B^{-1} \in \mathbb{R}^{n\times n}$ for A (also often called an *approximate inverse*) is a symmetric positive definite matrix, which satisfies the estimate

$$\delta_0(x, Bx) \le (x, Ax) \le \delta_1(x, Bx), \qquad \text{for all } x \in \mathbb{R}^n, \tag{1.6}$$

for some real positive constants δ_0 and δ_1. At this point we do not comment on the details of preconditioning, but refer to Section 1.1.5 for a brief survey on multigrid and multilevel preconditioning. We often assume, for the sake of simplicity, the somewhat simpler inequality

$$\left\|I - B^{-1}A\right\|_A \le \gamma, \tag{1.7}$$

for some real constant $\gamma \in [0, 1)$. Such an assumption is typically fulfilled for multigrid and domain decomposition preconditioners; for the best of them the constant γ is bounded away from 1 independently of the mesh size, see Section 1.1.5. Let us finally mention that in most cases the assumption (1.7) does not mean any restriction of the generality of the analysis, cf. Section 2.1.1 for details.

We emphasize that the preconditioner B^{-1} in our *matrix-free environment* is assumed to be only available as a matrix-vector-multiply function. Let us now include this operation into the following extended (in comparison with Section 1.1.2) set of admissible operations:

EXTENDED SET OF OPERATIONS FOR PRECONDITIONED EIGENSOLVERS:

1. Matrix-vector multiplications: $x \to Ax$ and $x \to Mx$,
2. Preconditioner-vector multiplications: $x \to B^{-1}x$,
3. Linear operations: $x + y$, λx,
4. Inner products: $(x, y) = x^T y$.

In a formal and preliminary way we derive a basic *preconditioned eigensolver* by premultiplying the residual $Ax - \lambda(x)Mx$ in the gradient method (1.5) by the preconditioner B^{-1}. This leads to the so-called *preconditioned gradient scheme*

$$x' = x - \omega B^{-1}(Ax - \lambda(x)Mx). \tag{1.8}$$

Once more, the scaling constant is to be determined appropriately. We first assume a properly scaled preconditioner so that $\omega = 1$. (Later in Chapter 6 we will analyze the preconditioned steepest descent scheme, where ω is constructed in a way so that the Rayleigh quotient $\lambda(x')$ is minimized.)

For a long time, without doubt, the area of preconditioned eigensolvers was dominated by Russian mathematicians. They were first suggested by Samokish [119] in 1958 and later by Petryshyn [110]. But there are exceptions, e.g. Ruhe [114] deals with convergent splittings for eigensolvers and Wachspress [132] proposes a combination of inverse iteration and multigrid iteration for the solution of the systems of linear equations. Some of the older works in this area are only available in Russian, which has not been beneficial to the popularization of such techniques in Western literature. Explicit convergence estimates proving grid independent convergence of preconditioned eigensolvers were given by Godunov et al. [46] and D'yakonov et al. [31, 35, 36]. A review of several results and the relevant literature is given in Chapter 9 of D'yakonov's monograph on optimization in solving elliptic problems [32]. Still an interesting source of ideas and proof techniques is Knyazev's PHD thesis, which first appeared in Russian [68] and whose main results are available in a translation [69].

While (1.8) is a vector scheme to compute the smallest eigenvalue together with an eigenvector, corresponding *subspace solvers* have been suggested and analyzed by Samokish [119], D'yakonov and Knyazev [33, 34], Meyer [89], D'yakonov [32], Bramble, Knyazev and Pasciak [15], Zhang, Golub and Law [149], Neymeyr [98] and others. We refer to Chapter 5 for a more detailed discussion including a brief review of the new analysis presented in [98]. A systematic classification of several preconditioned eigensolvers for symmetric positive definite eigenproblems and a survey on the literature till 1998 has been given by Knyazev [71]. See also [73] for a review on more recent developments.

It is important to note that the preconditioned eigensolver can be realized at optimal costs: Each step of the scheme (1.8) can be executed with small memory requirements, since only 3 vectors are required to hold x and the intermediate results Ax and Mx. The latter matrix-vector products are also needed to compute the Rayleigh quotient. We note that the necessity for storing 3 vectors (instead of 2) can be explained by the requirement for computing Ax and Mx only once per iteration of (1.8). The *total computational costs* behave like $\mathcal{O}(n)$, if $x \to B^{-1}x$ can be evaluated for optimal costs, too. In particular, 3 matrix-vector products are required (each one for A, M and B^{-1}), 2 inner products are needed to compute the Rayleigh quotient, and finally 2 operations $x + y$ and 1 operation λx are necessary to evaluate x'.

Some preconditioned eigensolvers have been discussed in the literature, which are based on similar ideas compared to that of the preconditioned gradient scheme: As an example let us consider Equation (1.8). For the case of exact preconditioning or $B = A$, we obtain the *descent direction* d, i.e. the direction in which the iterate x in (1.8) is corrected, as the solution of the linear system

$$Ad = Ax - \lambda(x)Mx. \tag{1.9}$$

In (1.9) the right-hand side is given by the residual belonging to x. Let us compare the latter equation with the *Davidson method* [27, 117] and the *Jacobi-Davidson method* [120–122]. In both cases the *correction equation* to be solved approximately for d is of the form

$$P(A - \lambda(x)M)Pd = Ax - \lambda(x)Mx. \tag{1.10}$$

Hence, Equations (1.9) and (1.10) differ in the operators on the left-hand sides,which are to be inverted approximately. For the Davidson method P equals the identity matrix, and (1.10) is solved approximately by using a *diagonal* preconditioner (roughly) approximating the inverse of $A - \lambda(x)M$. A further variant of the latter schemes is the Generalized Davidson method permitting also non-diagonal preconditioning, see Oliveira [103] and Ovtchinnikov [104] for recent results on its convergence. In contrast to this, for the Jacobi-Davidson scheme P is an orthogonal projector to the complement of the current eigenvector approximation. This projection provides the stabilization necessary for the more general scope of this solver to solve even unsymmetric and complex eigenvalue problems. For the analysis of the Jacobi-Davidson method applied to Hermitian positive definite matrix pencils see Notay [100, 101].

As a notable feature, convergence proofs for preconditioned eigensolvers are *difficult* and often *extremely technical*. Moreover, for some of the most effective and successful eigensolvers (e.g. the Jacobi-Davidson method of Sleijpen and van der Vorst [121] and the Locally Optimal Preconditioned Conjugate Gradient (LOBPCG) scheme as suggested by Knyazev

[70, 72]) no convergence theory is yet available—aside from trivial upper bounds which can be derived from related but less effective schemes.

There is a rule-of-thumb, sometimes formulated within the community of researchers on preconditioned eigensolvers, which says that it is easy to *suggest* a new eigensolver scheme (built on some combination of elements like orthogonal projections, Rayleigh-Ritz steps with respect to various subspaces, preconditioning strategies and so on), but that it is very hard to provide a sound convergence theory as well as to beat the performance of already known methods.

In this work we suggest a new theoretical framework for some class of preconditioned eigensolvers in which the iteration (1.8) is the most simple representative scheme. As pointed out in the next section, we prefer to derive preconditioned eigensolvers from a *preconditioned variant of inverse iteration*. Within this framework a new geometric interpretation is suggested which should convey a deeper understanding of preconditioning for the eigenproblem.

1.1.4 Justification of preconditioning for eigensolvers

It is by no means obvious that the replacement of the residual by the preconditioned residual, as done in (1.8) compared to (1.5), improves convergence. In other words, the demonstrated derivation of the preconditioned eigensolver (1.8) does not provide any explanation for its claimed efficiency. Of course, the convergence theory as contained in the cited papers (see Section 1.1.3) provides clear evidence for the superiority of preconditioned eigensolvers as, e.g., the convergence rates can be bounded away from 1 independently of the mesh size.

Furthermore, there is no consent in the literature on how to motivate and justify preconditioning for iterative eigensolvers. Hence, let us now systematically review the several approaches to the preconditioned eigensolver (1.8). Not all of them provide an "intuitive justification" for preconditioning, but we hope to give, at least with the two last approaches, a convincing explanation of why preconditioning makes a difference.

We first report on the approaches of D'yakonov [32], Knyazev [71] and Meyer [89], who all interpret preconditioned eigensolvers as modifications of the basic gradient-type eigensolver (1.5). Then we introduce a very different approach, which has been presented in [95, 96] by the author of this work. In these latter papers the root for deriving preconditioned eigensolvers is seen in the approximate solution of the system of linear equations associated with inverse iteration.

1. Taking the gradient with respect to the $(\cdot, B\cdot)$ inner product:

   D'yakonov [32, Section 9.4.1] points out that the search direction of (1.5) is the gradient of (1.3) in the Euclidean space $\mathbb{R}^n$, while the gradient, with respect to the $(\cdot, B\cdot)$ inner product, reads

$$\nabla_B \lambda(x) = \frac{2}{\|x\|_M^2} B^{-1}(Ax - \lambda(x)Mx). \tag{1.11}$$

Obviously, a correction of x in the direction $\nabla_B \lambda(x)$ defines the preconditioned gradient method. But this ansatz *does not offer a direct explanation for the question on why such preconditioning is advantageous*. Nevertheless, the convergence analysis by D'yakonov and Orekhov [36], D'yakonov and Knyazev [33] and D'yakonov [32] substantiates that the choice (1.11) hastens convergence—but here we would like to find further analytic and "intuitive" reasons why the B inner product is an advantageous selection.

The following fourth interpretation may shed light on (1.11): for exact preconditioning or $B = A$, the correction term $\nabla_A \lambda(x)$ together with $\omega = 1$ transforms (1.5) to (scaled) inverse iteration.

2. Applying the gradient method to the preconditioned eigenproblem:

   Knyazev [71, Section 4] compares the 2 eigensolvers

$$x' = x - \omega A^{-1}(Ax - \alpha Mx) \tag{1.12}$$

and

$$x' = x - \omega(Ax - \alpha Mx). \tag{1.13}$$

Therein the scalar α is an iteration parameter which can be identified with the Rayleigh quotient of x. In a second step the eigenproblem (1.2) is multiplied with the preconditioner leading to the preconditioned eigenproblem

$$B^{-1}Ax = \lambda B^{-1}Mx. \tag{1.14}$$

Obviously, neither the eigenvalues nor the eigenvectors of $(B^{-1}A, B^{-1}M)$ are changed in comparison to those of (A, M). The key point is that the two schemes (1.12) and (1.13) exhibit a very different behavior if applied to the preconditioned eigenproblem (1.14). Whereas replacing (A, M) by $(B^{-1}A, B^{-1}M)$ leaves (1.12) unchanged, the scheme (1.13) becomes a preconditioned one

$$x' = x - \omega B^{-1}(Ax - \alpha Mx). \tag{1.15}$$

It is argued in [71] that the eigenvectors of the iteration matrix $(A - \alpha M)$ in (1.13) are not the same as those of the original problem (1.2) and that this is the reason why preconditioning makes a difference.

Note that if α is identified with the Rayleigh quotient of x, then the substitution of (A, M) by $(B^{-1}A, B^{-1}M)$ in (1.3) has to be accompanied by a change of the Euclidean inner product to the inner product induced by B. Only such a simultaneous substitution can keep the Rayleigh quotient unchanged.

3. Preconditioning allowing to treat ill-conditioned eigenproblems:

   Meyer [89, Section 5.1] assumes (1.14) as a mesh eigenproblem for a self-adjoint and coercive elliptic partial differential operator. A reformulation of the preconditioned eigenproblem (1.14) as
   $$\mathcal{A}x = \kappa\mathcal{M}x \tag{1.16}$$
   is considered with $\mathcal{A} = B^{-1}M$, $\mathcal{M} = B^{-1}A$ and $\kappa = 1/\lambda$. In order to determine the *largest* eigenvalues $\kappa_i = 1/\lambda_i$ of (1.16), the convergence theory for the gradient method given by Longsine and McCormick [83] is employed. This is done with respect to the $(\cdot, B\cdot)$ inner product for which $\mathcal{A}$ and $\mathcal{M}$ are symmetric operators. While Meyer, throughout his work, discusses the case of subspace schemes only, we here restrict his arguments to a 1D subspace.

   There are two decisive factors proving the efficacy of the preconditioned scheme. The first point is that the gap number g
   $$g = \frac{1/\lambda_1 - 1/\lambda_2}{1/\lambda_1 - 1/\lambda_n}$$
   is of the order $\mathcal{O}(1)$ in the mesh parameter h, since e.g. $1/\lambda_n \simeq h^2$ for the Laplacian. The second point is based on the fact that the condition number $\kappa(B^{-1}A)$ for the best (multigrid) preconditioners is bounded independently of h. Having both quantities under control, the convergence theory of Longsine and McCormick [83] as well as Meyer [89] provides quantitative grid-independent estimates concerning the convergence of the eigenvalue approximations to the eigenvalues of problem (1.16), or (1.2) equivalently.

4. Preconditioned inverse iteration:

   Inverse iteration for computing the smallest eigenvalue, together with an eigenvector of (1.2), maps a given iterate x to the new iterate $\hat{x}$ by solving the linear system
   $$A\hat{x} = \kappa Mx, \tag{1.17}$$
   for some $\kappa \neq 0$. A normalization of $\hat{x}$ may follow in order to avoid numerical under-/ overflow. The choice of κ is immaterial for the convergence of inverse iteration. Hence, we may set $\kappa = \lambda$, the generalized Rayleigh quotient (1.3) of x. This choice of κ has the effect that $\hat{x} - x$ tends to 0 as $(x, \lambda(x))$ converges to an eigenpair and paves the way for the application of the preconditioner B^{-1}. Approximate solution of (1.17) results in the error propagation equation
   $$x' - \lambda A^{-1}Mx = (I - B^{-1}A)(x - \lambda A^{-1}Mx), \tag{1.18}$$
   in which the initial error $x - \lambda A^{-1}Mx$, i.e. the difference between the actual iterate x and the result of inverse iteration $\lambda A^{-1}Mx$, is multiplied by the error propagation matrix $I - B^{-1}A$ and results in the final error $x' - \lambda A^{-1}Mx$.

The crucial point now is that (1.18) can be rewritten in the following form, which does not contain the inverse of A

$$x' = x - B^{-1}(Ax - \lambda Mx). \tag{1.19}$$

This is the well-known preconditioned gradient scheme (1.8). Because of its derivation we prefer to call (1.19) *preconditioned inverse iteration*, abbreviated by PINVIT.

The *new derivation via the error propagation equation* (1.18) has given rise to a *new convergence analysis providing sharp non-asymptotic convergence estimates*. These results have been published in two papers by Neymeyr [95, 96]. On this basis somewhat simplified convergence estimates have been derived; they have appeared in a joint paper with Knyazev in [73].

The central results and conceptions of this analysis are summarized in Chapter 2; these results provide the basis for the further analysis contained in this work.

Formally, one can look at (1.18) as an inner/outer loop iteration: the inner loop uses preconditioning to solve (1.17) and the outer loop performs inverse iteration. We do not pursue this point of view since any inner loop solver can be replaced by a (more accurate) preconditioner providing the same result in a single step. Moreover, as will be highlighted in Chapter 3, exact preconditioning, i.e. $B = A$, is not the best choice for this scheme. Best preconditioning for the eigenvalue problem leads to the largest decrease of the Rayleigh quotient; in the most favorable case this would be a decrease to the smallest eigenvalue λ_1. Such an optimal decrease, which means one-step convergence to an eigenvector belonging to λ_1, may happen; cf. Lemma 3.2 and Corollary 4.8.

Remark 1.1. *Up to now we have discussed various ways of how to derive and motivate preconditioned eigensolvers. Beyond that, we have to state that there is no consent in the literature concerning the interpretation of preconditioned eigensolvers. It is not even clear how to determine the ideal preconditioner for an eigensolver. In the recent "Templates for the solution of algebraic eigenvalue problems", it is stated that "in general the matrix being preconditioned is nearly singular". In the light of the last interpretation (given above) this is clearly not the case since a preconditioner for A and not for $A - \lambda_1 M$ is considered! In general, the preconditioned eigensolvers analyzed in this work do not belong to the class of approximate shift-and-invert iterations, as we use preconditioners for the (nonsingular) discretization matrix. Nevertheless, it is very tempting to have a preconditioner approximating the inverse of $A - \lambda M$ in order to mimic the fast convergence of the Rayleigh quotient iteration. But as already pointed out in Sleijpen and van der Vorst [121] in the context of the Jacobi-Davidson scheme, the exact inverse of $A - \lambda_1 M$ would lead to no expansion of the search subspace.*

The same argument would apply for the preconditioned eigensolver (1.19). Inserting $B = (A - \lambda I)$ in the PINVIT scheme results in

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

The outcome would be an uncontrolled cancellation (in the case $\omega \approx 1$), or stationarity of the iteration otherwise. Such a phenomenon is often described in the literature and it is advised to apply a preconditioner which is not an "overly accurate" approximation of the shifted matrix $A - \lambda M$. Otherwise, as it is the case for the generalized Davidson scheme, one is faced with the difficulty that a further improvement of the quality of the preconditioner will destroy convergence increasingly.

1.1.5 Multigrid and multilevel preconditioning

What is the efficiency and computational complexity that can reasonably be expected for a multigrid eigensolver for mesh eigenproblems?

Any derivation of such an eigensolver should be guided by the highly efficient *multigrid solvers* for the numerical solution of *boundary value problems* for the class of operators under consideration. It is well known that such multigrid schemes can be understood as preconditioners. Therefore, in the light of preconditioned inverse iteration (compare its derivation in Section 1.1.4), it is clear that *any multigrid preconditioners as developed for the solution of a boundary value problem (for a self-adjoint and coercive elliptic partial differential operator) can be used as a "black box" for the solution of the corresponding eigenvalue problem.*

In Table 1.1 we schematically compare the treatment and solution of a *boundary value problem* with that of an *eigenvalue problem* for an elliptic partial differential operator. The first step toward the solution of both problems consists in their weak formulation. These are defined within proper Hilbert spaces. The second step is their discretization with respect to some finite-dimensional subspace, spanned by a basis of finite element functions, of the given Hilbert spaces. While the Galerkin discretization of the boundary problem results in the linear system $Ax = b$, the discretization of the variational eigenvalue problem leads to the generalized matrix eigenvalue problem (1.2). The main motivation for the present work was the observation of some unbalance between the set of possible solvers (and their convergence theory) for the boundary value problem and those for the eigenvalue problem:
On the one hand, numerical schemes for the iterative solution of linear systems of equations exploit or are based on the structure prescribed by the partial differential equation and its discretization. Examples are the classical schemes like the successive overrelaxation (SOR), the alternating direction iteration (ADI) as well as the more recent and very efficient multigrid and domain decomposition schemes. On the other hand, eigensolvers like inverse iteration, QR and Lanczos were developed from a point of view of numerical linear algebra—they do not profit from the structure of the spectral problem for (discretizations of) partial differential operators. Therefore, our goal is to develop an eigensolver, which should reach an efficiency

comparable to the typical efficiency of iterative solvers for the numerical solution of boundary value problems for this class of operators.

Obviously, the scheme of preconditioned inverse iteration *constitutes a link* between the *discrete eigenproblem* (1.2) and *multigrid preconditioners* for the solution of boundary value problems for self-adjoint and coercive elliptic boundary value problems.

Therefore, let us now give a brief survey on such highly efficient *multigrid schemes*, which allow to solve boundary value problems in the form of their corresponding linear systems of the dimension n with optimal complexity $\mathcal{O}(n)$. The key point is that they exploit the structure imprinted on the problem by the partial differential equation and its discretization.

Multigrid algorithms have been developed since the early 1960s. The starting point was Fedorenko's [41] two-grid scheme for the solution of the Poisson problem on the unit square. In 1964, Fedorenko extended this work to a W-cycle and gave a rigorous convergence proof showing grid-independent convergence for this problem. Names like Bachvalov, Astrakhantsev, Brandt, Bank and Dupont are connected with the further development of the theory of multigrid methods for H^2-regular boundary value problems [1, 4, 6, 18]. Hackbusch's works [51, 53, 54] from 1976 and the early 1980s mark a significant advance in the theory of multigrid and show that the convergence of the multigrid method (W-cycle) rests on an *approximation property* and a *smoothing property*. These properties have been proved for several customary smoothers and discretizations. These results [55] are often denoted as the *classical* multigrid theory. Similar results have been gained by Braess and Hackbusch for the V-cycle [12, 13]. But the above mentioned convergence theory is usually only applicable to uniform grids and H^2 regular problems.

The latter restrictions have been surmounted by the development and the analysis of so-called multi-level methods. The hierarchical basis preconditioner of Yserentant [144] belongs to that class. Quasi-optimal (i.e. $\mathcal{O}(n \log n)$) complexity has been shown without regularity assumptions and even for non-uniform triangulations. The major drawback of hierarchical basis multigrid methods is that the convergence in $\mathbb{R}^d$, $d \geq 3$, deteriorates, as the condition number increases exponentially in the number of grid levels. Let us finally mention that the multilevel BPX preconditioner of Bramble, Pasciak and Xu [17] features dimension-independent convergence behavior. The works on classical multigrid and multilevel methods have culminated in the papers of Bramble, Pasciak, Wang and Xu [16] and Xu [143], in which the classical multigrid and the new multilevel schemes have been presented within a unified theory of additive and multiplicative abstract subspace correction schemes. See also Yserentant [148] for a review of all these developments.

Let us summarize that the preconditioned eigensolver (1.19) inherits the mentioned favorable properties from multilevel preconditioning. Therefore, the importance of the preconditioned inverse iteration scheme can be explained from the fact that *optimal computational complexity* can even be guaranteed for *non-uniform grids* and *without any assumptions on the regularity*. Hence, preconditioned eigensolvers may appear as those optimal multigrid schemes searched for for a long time.

1.1.6 Multigrid solvers for elliptic eigenproblems

The aim of this section is to give a brief outline on alternative multigrid-based schemes for the numerical solution of the eigenvalue problem for elliptic partial differential operators, see also [75] for some short review on several applications.

The *benefit* of certain of these schemes (compared to the preconditioned eigensolver introduced so far) is their *wider range of applicability*, in a sense that they cannot only be applied to the eigenproblem for a self-adjoint and coercive elliptic partial differential operator. E.g., the *direct multigrid solver* of Hackbusch can be used for the general elliptic eigenvalue problem.

On the other hand, most of the multigrid eigensolvers mentioned below require elaborate programming techniques to write the program code. This is typically much more labor-intensive than using multigrid as a "black-box" as it is possible in the case of preconditioned eigensolvers (for which ready-to-use program codes can be taken from a library of multigrid solvers). So the major *drawback* of such alternative multigrid techniques is to be seen in the costs for writing multigrid eigensolver code and in the loss of flexibility, as various multigrid preconditioners can easily be applied to and tested within the setup of preconditioned eigensolvers.

Let us first specify those components which can usually be grafted on any such multigrid eigensolver:

- *Subspace extension*: Having designed an eigensolver to compute *only* the smallest eigenvalue together with an eigenvector, this vector scheme can be extended to a subspace algorithm in order to compute the invariant subspace belonging to some of the smallest eigenvalues. To this end the vector scheme is applied to each of the Ritz vectors spanning the actual subspace. By means of the Rayleigh-Ritz procedure the new Ritz values and Ritz vectors are computed. (An alternative but less stable strategy would be the application of the deflation technique.)

- *Nested iteration and adaptivity*: Mesh eigenvalue problems for partial differential operators are usually first given on a coarse grid. If the associated eigenproblem is relatively low dimensional, it can easily be solved by standard eigensolvers like QR. The coarse grid approximations are prolongated to some refined grid. By means of nested iteration the eigenvectors/values are computed on a sequence of refined grids by using a multigrid technique while the problem size increases considerably.

  Nested iteration can be combined with the concept of adaptivity: By generating adaptive grids, numerical approximation of the eigenvalues/vectors within a prescribed tolerance can often be gained with only a small portion of the necessary work when uniform grid refinement is employed. In order to construct an adaptive eigensolver one has to provide appropriate error estimators for the iteration error (to define a stopping criterion for the iterative solver on the actual grid) and for the discretization error (to control the mesh refinement), see [44, 97] and also Section 7.2.

A natural approach to the multigrid solution of the eigenvalue problem is to treat it as a nonlinear equation and to apply a nonlinear multigrid solver, e.g., the full approximation scheme (FAS), [20]. Another successful technique is the linearization of the discrete eigenproblem and to use multigrid as an inner solver. This linear solver is embedded in an outer iteration like the Rayleigh quotient iteration [87] or inverse iteration (with a shift) [7]. Whenever linear systems in indefinite matrices like $A - \sigma M$ (for some shift parameter σ) are to be solved, one is faced with the difficulty to define a termination criterion for the inner solver, cf. Section 1.2.4. Moreover, if σ is near to an eigenvalue of (A, M) the problem is almost singular. Whereas it is well known, due to the analysis of Peters and Wilkinson [109], that such a singularity does not destabilize inverse iteration with accurate solvers, it is a hard task to solve such equations approximately with *multigrid methods*. Therefore in the analysis of Bank [7] the shift parameter is bounded away from the eigenvalues of (A, M). One way to overcome this difficulty is the *direct multigrid approach* of Hackbusch [52, 55] whose central step is the solution of some correction equation within the orthogonal complement of the actual eigenvector approximation. For this reason the Hackbusch algorithm is intimately related with nested iteration in order to provide reliable coarse grid projections. For a given iterate x having the Rayleigh quotient $\lambda(x)$ the resulting two-grid method reads as follows:

$$\begin{aligned}
\tilde{x} &= Sx && \text{(smoothing step)}\\
d_c &= R(A - \lambda(x)M)\tilde{x} && \text{(coarse grid projection of the residual)}\\
d_c^{\perp} &= Q_c d_c && \text{(M-orthogonal projection)}\\
v_c &= (A_c - \lambda(x)M_c)^{-1} d_c^{\perp} && \text{(solution of correction equation)}\\
x' &= x - PQ_c v_c. && \text{(prolongation and correction)}
\end{aligned}$$

Therein, the index c denotes the coarse grid quantities. R is a restriction operator, P is a prolongation and Q_c is the orthogonal projection operator to the M-orthogonal complement of the actual eigenvector approximations. Finally, x' is the new eigenvector approximation. The coarse grid problem is no longer an eigenvalue problem but a singular equation, which can be solved recursively on the coarse grids. This is done by solving correction equations, each in the orthogonal complement of the previously computed coarse grid eigenvector approximations. For an application of this eigensolver to the square plate problem see [56, 61, 62].

As another successful multigrid eigensolver let us mention the Rayleigh quotient multigrid (RQMG) minimization technique suggested by Mandel and McCormick [22, 84]. In analogy to the (preconditioned) gradient scheme the idea is to consider the eigenvalue problem as an optimization problem for the Rayleigh quotient. In order to generate a sequence of iterates with a decreasing Rayleigh quotient, on each grid level k for each coordinate direction d_i^k (associated with the ith finite element function Ψ_i^k on level k) a coordinate relaxation scheme is applied, i.e. one computes the minimum

$$\lambda(x + \omega^* d_i^k) = \min_{\omega \in \mathbb{R}} \frac{(x + \omega d_i^k, A(x + \omega d_i^k))}{(x + \omega d_i^k, M(x + \omega d_i^k))}. \tag{1.20}$$

The decisive point for an efficient implementation of RQMG is the appropriate choice of the restriction operators, which enable an *exact* representation of the Rayleigh quotient of the final level on all coarser grids. The convergence theory for RQMG, also showing its grid independent convergence, is given in [22, 88]. The RQMG method has been integrated into an adaptive 2D Helmholtz eigensolver used for designing integrated optical chips [28, 44]. A generalization to a subspace scheme for non-self adjoint elliptic operators has been devised by Deuflhard et al. [44].

Other variants of multigrid eigensolvers for elliptic differential operators working with optimal or quasi-optimal computational complexity have been invented, e.g. for a selection of those by Russian (co)authors see Astrakhantsev [2], Strakhovskaya and Fedorenko [125, 126] as well as Chan and Sharapov [23].

A fair comparison of the efficiency of different multigrid schemes is a difficult task as the methods often have a somewhat different scope and, therefore, their implementations are not optimized to treat a common test problem in the most efficient way. Nevertheless, the numerical tests in [82] give no indication on the superiority of one of the tested schemes.

1.2 A hierarchy of preconditioned eigensolvers

In this section we would like to propose a *new framework for preconditioned eigensolvers*. Within this framework we recover not only the scheme of preconditioned inverse iteration (see interpretation 4 in Section 1.1.4) as the most simple representative, but also the well-known *Preconditioned steepest descent* method and the *Locally optimal Block Preconditioned Conjugate Gradient* (LOBPCG) scheme [5].

Our aim is to show how these eigensolvers (together with some of their variants) can systematically be derived as iterations approximating some modification of *subspace iteration* with an improved computation of the Rayleigh-Ritz approximations.

Remark 1.2 (Reduction to the standard eigenproblem). *For the sake of clear and succinct representation we consider in the following only the standard eigenvalue problem, i.e. we formally set $M = I$ in (1.2). To justify this simplification, one exploits the fact that M is a symmetric positive definite matrix and thus defines the inner product $(\cdot, M\cdot)$. With respect to this inner product $M^{-1}A$ and $M^{-1}B$ are symmetric positive definite matrices, too. Now changing at once*

$$(\cdot,\cdot) \to (\cdot, M\cdot), \quad A \to M^{-1}A, \quad B \to M^{-1}B,$$

transforms the eigensolver (1.8) for the generalized eigenproblem applied to (A, M) to that for the standard eigenproblem for A, i.e.

$$x' = x - \omega B^{-1}(Ax - \lambda(x)x), \quad \textit{with} \quad \lambda(x) = \frac{(x, Ax)}{(x, x)}.$$

Beyond that, this transformation does not change the form of the quality condition (1.7) on the preconditioner, see also [97] and Theorem 4 in [73].

1.2.1 Subspace iteration

Let us start with a description of subspace iteration, which is the straightforward generalization of the power method or inverse iteration. Subspace iteration schemes were mainly developed in the 1960s and 1970s, see Chatelin [24] or Parlett [107]. In this work we are only interested in determining several of the *smallest* eigenvalues of A together with the invariant subspace. Hence, we will only discuss subspace iteration for A^{-1} (instead of that for A), where A is assumed to be a symmetric and positive definite matrix.

Subspace iteration is based on the following conception: Given a subspace $\mathcal{S}$ of the $\mathbb{R}^n$ having the dimension s, then repeated application of A^{-1} defines the Krylov space

$$\mathcal{K}(\mathcal{S}) = \mathrm{span}\{\mathcal{S}, A^{-1}\mathcal{S}, A^{-2}\mathcal{S}, \ldots\}. \tag{1.21}$$

Let $\mathcal{S}$ be spanned by the columns of $[z_1, \ldots, z_s] \in \mathbb{R}^{n\times s}$. It is well known that each column of $A^{-j}\mathcal{S}$, for $j = 1, 2, \ldots$, will converge to an eigenvector associated with the smallest eigenvalue as long as none of the $z_1, \ldots, z_s$ is orthogonal to the invariant subspace to λ_1. Since each $A^{-j}z_i$, $i = 1, \ldots, s$, will converge to that invariant subspace, orthogonality between the $A^{-j}z_i$ gets more and more lost, in particular if λ_1 is a non-degenerate eigenvalue. Then $A^{-j}[z_1, \ldots, z_s]$ will become a very poor basis for the "good" subspace $A^{-j}\mathcal{S}$. The idea of subspace iteration consists in iteratively constructing an orthonormal basis V_j of $A^{-j}\mathcal{S}$ in the following way:

Algorithm 1.3 (Subspace iteration, INVIT(1,s)).

i. Compute an orthonormal basis $V_1 \in \mathbb{R}^{n\times s}$ of $\mathcal{S}$.

ii. For $j \geq 1$ solve $AU_{j+1} = V_j$ for U_{j+1} and determine an orthonormal V_{j+1} with the same column space as U_{j+1}.

Orthonormalization can either be carried out with the Gram-Schmidt or the Rayleigh-Ritz procedure [107] (where a clever implementation is available which avoids extra matrix-vector multiplications in order to form the Rayleigh-Ritz projection, see Parlett [107, Section 14.2] and Contribution II/9 in Wilkinson and Reinsch [142]). Here, we prefer V_{j+1} to be computed by means of the Rayleigh-Ritz procedure so that the columns of V_{j+1} consist of Ritz vectors which are in several senses the optimal eigenvector approximations.

One might look upon subspace iteration as an outdated algorithm since more efficient Krylov subspace algorithms have been developed within the last two decades. Parlett [107] describes two conditions under which subspace iteration still appears attractive.

1. For very large problems the computer storage may be so limited that one can only hold a fixed small number of vectors. Under those circumstances one is forced to discard previous vectors from the Krylov space. Subspace iteration is the resulting method.

2. Whenever the relative gap between the wanted eigenvalues and the remaining ones is sufficiently large, subspace iteration may converge in only a few steps. In these situations the superior properties of methods working on full Krylov space (like Lanczos) are given no chance.

The first argument fully applies to the problems we intend to solve (while the second point is most striking if a shifted iteration operator like $A - \sigma I$ is applied). Discretizations of partial differential operators often lead to extremely large problems and only a small number of vectors, say less then 10, can be stored.

The pleasant feature of subspace iteration, namely that A is neither modified nor needs to be known explicitly, applies, for instance, to finite element methods. Finite element codes typically provide only routines for computing Ax: Either A is stored in some sparse matrix format or the matrix-vector product Ax is evaluated by a local compilation procedure. But in order to implement subspace iteration for A^{-1}, we need a linear system solver. In the context of discretized partial differential operators, these solvers are built on iterative methods (like multigrid or domain decomposition) and their application can be represented by some preconditioner, see Sections 1.1.3 and 1.1.5.

In order to prepare the ground for introducing *preconditioned subspace iteration*, we define some variant of subspace iteration in which the Rayleigh-Ritz procedure is applied to some enlarged subspace. The simple idea is to hasten convergence by adding to the subspace U_{j+1} (to which Rayleigh-Ritz [107] is applied in Algorithm 1.3) a number of $k-1$ of the previous subspaces $V_j, \ldots, V_{j-k+2}$. Then Rayleigh-Ritz (RR) works on a subspace, having the dimension ks, and V_{j+1} is formed by the s Ritz vectors corresponding to the s smallest Ritz values. The Courant-Fischer principle [107] guarantees that the s smallest Ritz values computed in this way are smaller than those computed by standard subspace iteration.

Algorithm 1.4 (Subspace iteration with improved RR projection, INVIT(k,s)).

i. Initialization: Given a subspace $\mathcal{S}$ compute orthonormal bases of $V_1, \ldots, V_{k-1} \in \mathbb{R}^{n \times s}$ of the $k-1$ subspaces

$$\mathcal{S}, A^{-1}\mathcal{S}, \ldots, A^{-(k-2)}\mathcal{S}.$$

ii. Iteration: For $j = k-1, k, k+1, \ldots$ solve the linear system

$$AU_{j+1} = V_j \tag{1.22}$$

for U_{j+1} and apply Rayleigh-Ritz to the system of k subspaces

$$[V_{j-k+2}, \ldots, V_j, U_{j+1}] \in \mathbb{R}^{n \times ks}.$$

Then let $V_{j+1} = [v_1, \ldots, v_s]$, where v_i denote the s orthonormal Ritz vectors associated with the s smallest Ritz values.

There is no necessity to define a new Krylov space for Algorithm 1.4, since the iterates of the improved subspace iteration are enclosed in $\mathcal{K}(\mathcal{S})$.

1.2.2 Preconditioned subspace iteration

We have introduced two variants of subspace iteration and we will now present its preconditioned variants. First let us introduce some naming allowing their classification. Algorithm 1.3 is called INVIT(1,s), expressing that inverse iteration (INVIT) is applied to an s-dimensional initial subspace $\mathcal{S}$ and that Rayleigh-Ritz only works on the single space $\operatorname{span}(U_{j+1})$. In contrast to this, Algorithm 1.4 is denoted INVIT(k,s) since Rayleigh-Ritz works on the k previous iterates V_j of the Krylov space (1.21). To be consistent with the usual nomenclature we abbreviate INVIT(1,1) by INVIT.

Let us now solve Equation (1.22) approximately by using preconditioning. As pointed out in [98], a necessary prerequisite for applying preconditioning is to substitute (1.22) by

$$AU_{j+1} = V_j \Theta_j, \tag{1.23}$$

where $\Theta_j = \operatorname{diag}(\theta_1, \ldots, \theta_s) = V_j^T A V_j$ is the diagonal matrix of the actual Ritz values giving rise to some column scaling of V_j. Observe that the choice of these (nonzero) scaling constants is immaterial for inverse iteration, since the convergence measures (like the Rayleigh quotient or the angle enclosed with invariant subspaces of A) do not depend on scaling. But scaling has the positive effect that the residual matrix

$$R := AV_j - V_j \Theta_j$$

converges to the zero matrix as the subspace spanned by V_j converges to an invariant subspace of A. This paves the way for the application of the preconditioner B^{-1}, which is assumed to satisfy (1.7). We get the update formula

$$\tilde{U}_{j+1} = V_j - B^{-1}(AV_j - V_j \Theta_j), \tag{1.24}$$

where $\tilde{U}_{j+1}$ approximates the solution U_{j+1} of (1.22).

Let us define the scheme PINVIT(k,s) of preconditioned (modified) subspace iteration.

Algorithm 1.5 (Preconditioned subspace iteration, PINVIT(k,s)).

i. *Compute an orthonormal basis of* V_1 *of* $\mathcal{S}$ *and additionally* $k-2$ *orthonormal bases* $V_2, \ldots, V_{k-1} \in \mathbb{R}^{n \times s}$*, for instance by using PINVIT(j,s) for* $j = 2, \ldots, k-1$.

ii. *For* $j \geq k-1$ *let*

$$\tilde{U}_{j+1} := V_j - B^{-1}(AV_j - V_j \Theta_j) \tag{1.25}$$

and apply Rayleigh-Ritz to

$$[V_{j-k+2}, \ldots, V_j, \tilde{U}_{j+1}] \in \mathbb{R}^{n \times ks}.$$

Then $V_{j+1} = [v_1, \ldots, v_s]$ *is composed of the* s *Ritz vectors belonging to the* s *smallest Ritz values.*

	(Preconditioned) subspace iteration with RR	Vector iteration scheme $\dim(\mathcal{S}) = 1$	No Rayleigh-Ritz $k = 1$
$\gamma = 0$	INVIT(k,s)	INVIT(k)	INVIT
$\gamma \geq 0$	PINVIT(k,s)	PINVIT(k)	PINVIT

Table 1.2: *Classification of (preconditioned) subspace iteration schemes*

Remark 1.6. *If $k > 1$, we simply write (instead of (1.25))*

$$\tilde{U}_{j+1} := B^{-1}(AV_j - V_j\Theta_j),$$

since then V_j is contained in the subspace to which Rayleigh-Ritz is applied.

First observe that the iterates of PINVIT(k,s) are not contained in the Krylov space (1.21). See Knyazev [72] for the definition of a generalized Krylov space $\hat{\mathcal{K}}$ based on polynomials of two independent variables and which contains the iterates of PINVIT(k,s). Note that for a more and more accurate preconditioner the acute angle $\measuredangle(\hat{\mathcal{K}}, \mathcal{K})$ between these Krylov spaces tends to 0. Accordingly, PINVIT(k,s) reduces to INVIT(k,s) for accurate preconditioning, which means $B = A$ or $\gamma = 0$ with respect to (1.7).

For a major part in this work we will discuss the preconditioned *vector* schemes PINVIT(k,1) which we abbreviate, for simplicity, by PINVIT(k). If additionally no Rayleigh-Ritz is applied, i.e. $k = 1$, we will simply write PINVIT instead of PINVIT(1), in accordance with the usual notation used in [95, 96]. Table 1.2 summarizes the notation.

In order to code the eigensolvers PINVIT(k,s) we first have to provide a routine computing the product Ax for given x, and secondly, a procedure which gives back $B^{-1}x$, where B^{-1} is an approximate inverse (or preconditioner) of A. Finally, the application of the Rayleigh-Ritz procedure requires the computation of additional matrix-vector products with A and several inner products.

Let us now state in Lemma 1.7 that PINVIT(m,s) converges stepwise faster than PINVIT(k,s) if $m > k$. Admittedly, its proof is a simple consequence of the min-max principles. Nevertheless, its statement is worthwhile, since we will present *sharp* estimates for PINVIT(1,s) in this work. These bounds can serve as trivial upper estimates for PINVIT(k,s), $k \geq 2$. It is important to note, that these estimates, at least for $k > 2$, are the best non-asymptotic estimates so far available. See also Chapter 6 for PINVIT(2) convergence estimates.

Lemma 1.7. *If $m, s \in \mathbb{N}$, then PINVIT(m,s) for $m \geq k$ does not converge more slowly than PINVIT(k,s) in the following sense: For the s smallest Ritz values $\theta_i^{(m)}$ computed from the subspace*

$$[V_{j-m+2}, \ldots, V_j, \tilde{U}_{j+1}] \in \mathbb{R}^{n \times ms}$$

and the s smallest Ritz values $\theta_i^{(k)}$ determined from $[V_{j-k+2},\dots,V_j,\tilde{U}_{j+1}] \in \mathbb{R}^{n\times ks}$ it holds that

$$\theta_i^{(m)} \leq \theta_i^{(k)}, \qquad i = 1,\dots,s.$$

□

1.2.3 Subspace iteration and gradient type schemes

Having introduced the shorthand notation PINVIT(k,s), let us now recover those eigensolvers which are customarily subsumed under these synonyms. For $k = 1, 2, 3$ the following names are used in the literature:

PINVIT	Preconditioned gradient method/Preconditioned inverse iteration,
PINVIT(2)	Preconditioned steepest descent,
PINVIT(2,s)	Preconditioned block steepest descent,
PINVIT(3)	LOPCG, Locally Orthogonal Preconditioned Conjugate Gradient,
PINVIT(3,s)	LOBPCG, Locally Orthogonal Block PCG.

Our aim is now to make clear that the preconditioned eigensolvers PINVIT(k,s) are in no sense “close” to gradient type eigensolvers. Instead, by the derivation of these schemes in the last section, the *preconditioned eigensolvers approximate subspace iteration for A^{-1}*, while *gradient type eigensolvers are associated with subspace iteration for A.*

To explore these relations, we insert $B = I$, an extremely poor choice of the preconditioner, into the schemes PINVIT(1) and PINVIT(2):

1. PINVIT for $B = I$ reduces to

$$(x,\lambda) \quad \longrightarrow \quad (x' := x - (Ax - \lambda x), \lambda(x')), \tag{1.26}$$

which is simply the gradient method (1.5) for $\omega = 1$.

2. In the same way the choice $B = I$ leads from PINVIT(2) to

$$(x,\lambda) \quad \longrightarrow \quad (v_1(V), \theta_1(V)) \quad \text{with} \quad V = [x, Ax], \tag{1.27}$$

where v_1 is the Ritz vector to the smallest Ritz value θ_1 of V. We rewrite this as

$$(x,\lambda) \quad \longrightarrow \quad (x' := x - \omega(Ax - \lambda x), \lambda' := \lambda(x')), \tag{1.28}$$

where ω minimizes the Rayleigh quotient of x'. This makes it understandable why (1.27) and (1.28) are called *steepest descent* for the Rayleigh quotient; cf. Section 1.1.2 and the last column of Table 1.3.

Therefore, for $B = I$ both schemes act in the Krylov space

$$\tilde{\mathcal{K}} = \operatorname{span}\{\mathcal{S}, A\mathcal{S}, A^2\mathcal{S}, \ldots\},$$

associated with *subspace iteration for* A. We emphasize that the choice $B = I$ is way out from our quality conditions (1.6) and (1.7) on the preconditioner. Inserting $B = I$ in (1.7) results in

$$\|I - A\|_A = \|I - A\|_2 \approx \lambda_{\max}(A).$$

Since the latter quantity behaves like h^{-2} for the discrete Laplacian Δ_h, it holds $\|I - A\|_2 \gg 1$ in any interesting case. Hence, the choice $B = I$ does not meet the quality conditions on the admissible preconditioners. In other words, there is no admissible preconditioner making PINVIT(1) or PINVIT(2) reduce to the gradient method (1.26) or steepest descent (1.27). For this reason we prefer to consider our class of preconditioned eigensolvers as preconditioned inverse iteration and most often avoid to call them preconditioned gradient schemes.

In Table 1.3 we summarize the mentioned relations between PINVIT(k,1) and INVIT(k,1) for $k = 1, 2, 3$ as well as the corresponding gradient schemes; the shorthand notation $x^{[-m]}$ is used to denote older iterates.

1.2.4 Generalized inexact solvers

The idea underlying the derivation of the preconditioned subspace schemes PINVIT(k,s) in Section 1.2.2 can be generalized to an inner-outer loop structure solver, realizing some form of *inexact inverse iteration*. The outer loop is based on a shift-and-invert transformation, i.e. inverse iteration with a shift or the Rayleigh quotient iteration. Multigrid preconditioning of indefinite matrices is a delicate task [99, 130, 145, 146]; we do not analyze eigensolvers based on preconditioning for indefinite matrices in the present work. The inner loop can be realized by any (approximate) linear solver, possibly a Krylov subspace solver can be used.

Iterative eigensolvers based on this idea have been suggested by Smit and Paardekooper [123] and Golub and Ye [49] and Lai, Lin and Lin [77]. In all these works a stopping condition for the inner iteration is constructed, guaranteeing that the outer loop converges at least linearly. Thus the threshold parameter is the central control parameter for the rate of convergence. The convergence analysis is based on a decomposition of the iteration vector in the first eigenvector and its orthogonal complement. The geometric way of decomposing the iterates in [123] resembles, in some sense, the one used in the present work.

The main difference between [49, 77, 123] and this work is that the stopping condition in the cited papers is an *a posteriori* criterion. In our setup the quality condition (1.7) on the preconditioner should be considered as an *a priori* criterion. Therefore, we do not see any necessity to describe PINVIT as an inner-outer iteration scheme. In other words, we do not consider multiple inner solution steps for two reasons: On the one hand, any multiple-step inner solver can be substituted by a one-step solver based on a more accurate preconditioner. On the other hand, we emphasize that an accurate linear solver does not pay out in general.

k	Subspace Iteration with improved RR projection	Preconditioned Subspace Iteration with improved RR projection	Exact preconditioning $B = A$ or $\gamma = 0$	No preconditioning $B = I$
$k = 1$	INVIT $x \to x' = A^{-1}x$ $\lambda \to \lambda(A^{-1}x)$	PINVIT $x \to x' = x - B^{-1}(Ax - \lambda x)$ $\lambda \to \lambda(x')$	scaled INVIT $x \to x' = \lambda A^{-1}x$ $\lambda \to \lambda(A^{-1}x)$	Gradient method $x \to x' = x - (Ax - \lambda x)$ $\lambda \to \lambda(x')$
$k = 2$	INVIT(2) $(x, \lambda) \to (v_1(V), \theta_1)$ $V = [x, A^{-1}x]$	PINVIT(2), also called Preconditioned Steepest Descent $(x, \lambda) \to (v_1(V), \theta_1)$ $V = [x, B^{-1}(Ax - \lambda x)]$	INVIT(2) $(x, \lambda) \to (v_1(V), \theta_1)$ $V = [x, \lambda A^{-1}x]$	Steepest descent $(x, \lambda) \to (v_1(V), \theta_1)$ $V = [x, Ax - \lambda x]$
$k = 3$	INVIT(3) $(x, \lambda) \to (v_1(V), \theta_1)$ $V = [x^{[-1]}, x, A^{-1}x]$	PINVIT(3), also called LOBPCG for $s \geq 1$, [72]. $(x, \lambda) \to (v_1(V), \theta_1)$ $V = [x^{[-1]}, x, B^{-1}(Ax - \lambda x)]$	INVIT(3) $(x, \lambda) \to (v_1(V), \theta_1)$ $V = [x^{[-1]}, x, \lambda A^{-1}x]$	“Generalized gradient type method” $(x, \lambda) \to (v_1(V), \theta_1)$ $V = [x^{[-1]}, x, Ax - \lambda x]$
$\vdots$ $\vdots$	$\vdots$ $\vdots$	$\vdots$ $\vdots$	$\vdots$ $\vdots$	$\vdots$ $\vdots$
$k \in \mathbb{N}$	INVIT(k) $(x, \lambda) \to (v_1(V), \theta_1)$ $V = [x^{[2-k]}, \ldots, x, A^{-1}x]$	PINVIT(k) $(x, \lambda) \to (v_1(V), \theta_1)$ $V = [x^{[2-k]}, \ldots, x, B^{-1}(Ax - \lambda x)]$	INVIT(k) $(x, \lambda) \to (v_1(V), \theta_1)$ $V = [x^{[2-k]}, \ldots, x, A^{-1}x]$	$(x, \lambda) \to (v_1(V), \theta_1)$ $V = [x^{[2-k]}, \ldots, x, Ax - \lambda x]$

Table 1.3: The hierarchy of preconditioned eigensolvers PINVIT(k,s), here $s = 1$.

Fastest convergence of the preconditioned eigensolvers PINVIT(k,1) is not guaranteed by exact preconditioning! For the analysis proving the latter proposition, we refer to Chapter 3 and to the estimates presented in Chapter 4. There we can construct specific preconditioners, which are poor for the solution of linear systems in A, but which, at the same time, may make one-step convergence of the eigensolver possible.

So far no decisive answer is available concerning the question of whether positive definite or indefinite preconditioners (e.g. approximating the Rayleigh quotient iteration) lead to more efficient algorithms, cf. Knyazev [71]. This opens the question of how accurately the associated equations are to be solved in order to achieve a reasonable speed of convergence [49, 77, 123]. In any way the cubic convergence of the Rayleigh quotient iteration cannot be transferred to the preconditioned multigrid case.

1.3 An application: Electronic structure theory

As a challenging area of application of multigrid preconditioned eigensolvers, we would like to highlight the *molecular electronic structure theory* as a subfield of quantum theory, which was founded by Schrödinger and Heisenberg between 1924 and 1926. As early as in 1929 Dirac formulated [29]: “The physical laws necessary for the mathematical theory of a large part of physics and the whole chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble.” These difficulties have not been surmounted yet: Even under several simplifications of the governing principles/equations, the numerical treatment of such problems is often extremely expensive.

The neglect of time-dependent interactions in (parts of) the molecular electronic structure theory [127] allows us to restrict ourselves to the *time-independent non-relativistic Schrödinger equation*

$$\mathcal{H}\Psi = E\Psi. \tag{1.29}$$

For an isolated N-electron atomic or molecular system in the Born-Oppenheimer approximation (which treats the nuclei as stationary sources of an electric field rather than as true particles), the Hamiltonian operator in atomic units reads

$$\mathcal{H} = -\sum_{i=1}^{N} \frac{1}{2}\Delta(r_i) + \sum_{i=1}^{N} v(r_i) + \sum_{i<j}^{N} \frac{1}{r_{ij}}.$$

Therein, r_i is the space coordinate of the ith electron, and r_{ij} ($r_{i\alpha}$) are its distances to the electron j (the nucleus α). In addition, $v(r_i) = \sum_\alpha Z_\alpha / r_{i\alpha}$ denotes the potential acting on the electron i, which is induced by the nuclei α with the charges Z_α.

Hence the Schrödinger equation is a differential eigenvalue problem for the wave function

$$\Psi = \Psi(r_1, s_1, r_2, s_2, \dots, r_N, s_N), \tag{1.30}$$

which depends on the r_i and spin coordinates s_i, $i = 1, \ldots, N$. Each eigenvalue E is considered as the energy related to the corresponding eigenstate Ψ.

The *Hartree-Fock approximation* is the most important *orbital method* for solving the molecular Schrödinger equation. To this end, the Hartree-Fock equations are discretized with respect to a linear expansion of the unknown molecular orbitals. The resulting set of equations for the expansion coefficients are called the Roothaan equations—they represent a nonlinear generalized eigenvalue problem, which is solved iteratively by the *self-consistent field* procedure. The resulting Hartree-Fock wave functions provide an approximate solution of (1.29) for closed-shell molecules.

As motivated by physical/chemical considerations, the wave function is usually expanded in Slater or contracted Gaussian functions, which are essentially the one-electron wave functions of the hydrogen atom. By using such a basis, the overall algorithm allows us to treat even relatively large (organic) molecules. The major drawback of the Hartree-Fock-Roothaan approach is its computational complexity, which increases as $\mathcal{O}(M^4)$ in the number M of basis functions. This unfavorable scaling is a consequence of the M^4 two-electron integrals, which are needed to form the discrete Fock operator. Considerable effort has been made to reduce this order and several more elaborate schemes have been devised in the last decade [102].

From the view of numerical analysis, it is also tempting to expand the wave function in terms of finite element functions. This can be done successfully (in spite of their bounded support) for relatively small molecules, like the hydrogen molecule, for which the molecular symmetry can be exploited in an advantageous way [134]. The major difficulty concerning the application of the finite element method is that the wave function, as given by (1.30), depends on the (large) number of $3N$ spatial coordinates.

A promising way to reduce this high-dimensional problem is the *density functional theory* [108], which states that the ground state wave function is determined by the electron density $\rho(x)$ in such a way that one can work without loss of rigor with the electron density $\rho(r)$. Since the electron density depends on only 3 spatial variables, one can classify computational schemes derived from the density functional theory as *real-space* methods. Within this theory one has to solve the Kohn-Sham equations (1.31), which represent a low dimensional approximation to ab-initio quantum chemistry, i.e, one has to determine the N eigenfunctions (representing the N electrons) belonging to the smallest eigenvalues ϵ_i of the problem

$$\begin{aligned} (-\Delta + \hat{V})\psi_i(x) &= \epsilon_i \psi_i(x), \qquad && x \in \Omega, \\ \psi_i(x) &= 0, && x \in \partial\Omega, \end{aligned} \tag{1.31}$$

for a bounded domain $\Omega = [0, L]^3$, $L > 0$. Therein, $\hat{V}$ is a symmetric non-local operator, which depends on the positions of the nuclei and on global integrals involving the (unknown) eigenfunctions. For problems of electronic structure theory, the spectrum of the operator $-\Delta + \hat{V}$ is bounded from below and the (possibly multiple) smallest eigenvalues of interest are negative.

The central tasks are to solve not only an eigenvalue problem for the Laplacian but also to determine the global integrals by solving the Poisson equation. Therefore, the outlined situation fits perfectly into the setup of multigrid preconditioned eigensolvers as presented in this work. (With a proper shift one can transform (1.31) into a coercive problem.) Both solvers scale like $\mathcal{O}(N)$ and the well-understood multigrid preconditioners for the Laplacian can be used.

Within the density functional theory and by using the outlined approximations and techniques, very recent results show that the entire problem can be solved numerically with only $\mathcal{O}(N \log N)$, or even better $\mathcal{O}(N)$, operations [9, 19, 40, 133].

1.4 Model analysis of inverse iteration

We conclude Chapter 1 with a model analysis of *inverse iteration* (INVIT) in which we highlight inverse iteration as a *descent method for the Rayleigh quotient*.

Obviously, we do not claim to present new results for the well-understood and simple scheme of inverse iteration [24, 107, 136]. The aim of this section is to point out an unusual representation of the convergence results for INVIT as given by Theorem 1.8 and Corollary 1.10. This includes the introduction of some convergence measures in terms of the $\Delta_{p,q}$ ratios, see Equation (1.37). Additionally, we would like to introduce the *Lagrange multiplier method* as a valuable tool for analyzing eigensolvers, see the proof of Theorem 1.8. In the following chapters of this work the Lagrange multiplier method turns out very useful for the PINVIT(k) analysis. In some sense the analysis of INVIT given here has some model character as its convergence estimates can be derived in an easy way. In contrast to this, it will cost us much more effort to derive comparable convergence estimates based on similar measures for the (improved) techniques PINVIT(1), INVIT(2) and PINVIT(2).

Inverse iteration is a simple vector iteration for computing the eigenvalue with the smallest absolute value together with an eigenvector of a regular symmetric matrix. Inverse iteration goes back to Wielandt [136], see also [63] for remarks on its history. Here we consider inverse iteration without a shift. Given a regular and symmetric $A \in \mathbb{R}^{n\times n}$ with real eigenvalues $0 < |\lambda_1| < |\lambda_2| \leq \ldots \leq |\lambda_n|$ (we denote the corresponding normed eigenvectors by $x_1, \ldots, x_n$), then the "standard" convergence analysis is based on an eigenvector expansion of an initial vector $x = \sum_{i=1}^n \alpha_i x_i$. We assume $\alpha_1 \neq 0$. Hence, in

$$A^{-1}x = \sum_{i=1}^{n} \frac{\alpha_i}{\lambda_i} x_i$$

all components with indexes $i \neq 1$ are damped out relatively and, therefore, inverse iteration converges linearly to the smallest eigenpair (x_1, λ_1). Equivalently, one can employ a two dimensional analysis in the plane containing the actual iterate and the wanted eigenvector. Once more, all components orthogonal to the wanted eigenvector are damped out. See Chatelin [24], Golub and van Loan [48], Parlett [107] and Stoer and Bulirsch [124].

Here we pursue an alternative approach to a convergence analysis of inverse iteration which is restricted to symmetric and positive definite matrices. For this class of matrices inverse iteration can be seen as a *descent method for the Rayleigh quotient*. To be precise, the iterates form a sequence of vectors with a *monotone decreasing* Rayleigh quotient. The Rayleigh quotients are guaranteed to converge to *an* eigenvalue of A. Under the assumptions of the following theorem this limit is not necessarily the smallest eigenvalue λ_1 but possibly a larger eigenvalue; cf. Remark 1.9. Nevertheless, the sequence of vector-iterates is guaranteed to converge to a corresponding eigenvector of A.

In the next theorem, we present sharp bounds from above and below for the decrease of the Rayleigh quotients of the INVIT iterates.

Theorem 1.8. *Consider a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ having the eigenpairs (x_i, λ_i) and $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_n$. For any nonzero initial vector $x^{(0)}$ the iterates $x^{(k)}$ of inverse iteration*

$$A\tilde{x}^{(k+1)} = x^{(k)}, \qquad x^{(k+1)} = \frac{\tilde{x}^{(k+1)}}{\|\tilde{x}^{(k+1)}\|}, \qquad k = 0, 1, \ldots,$$

converge to an eigenvector of A, and the $\lambda(x^{(k)})$ converge to the corresponding eigenvalue. If the Rayleigh quotient $\lambda = \lambda(x^{(k)})$ of the actual iterate satisfies $\lambda \in (\lambda_i, \lambda_{i+1})$ for some i with $1 \leq i < n$, then it holds a sharp estimate from below and above for the Rayleigh quotient of $x^{(k+1)}$

$$B(\lambda_1, \lambda_n, \lambda) \leq \lambda(x^{(k+1)}) \leq B(\lambda_i, \lambda_{i+1}, \lambda) < \lambda, \tag{1.32}$$

where

$$B(\lambda_p, \lambda_q, \lambda) = \left(\lambda_p^{-1} + \lambda_q^{-1} - (\lambda_p + \lambda_q - \lambda)^{-1}\right)^{-1}. \tag{1.33}$$

Proof. Let $U^T AU = \Lambda$, where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, and let x be the actual iterate. Then $v = U^T x$ is the coefficient vector with respect to the basis of eigenvectors. We obtain for the Rayleigh quotient of $A^{-1}x$

$$\lambda(A^{-1}x) = \frac{(v, \Lambda^{-1}v)}{(v, \Lambda^{-2}v)} =: \lambda'.$$

In order to determine the extrema of λ' with respect to all $x \in \mathbb{R}^n$ with $(x, x) = 1$ and $(x, Ax) = \lambda$ we apply the method of Lagrange multipliers. The Lagrange function $L(v, \mu, \nu)$ reads

$$L(v, \mu, \nu) = \frac{(v, \Lambda^{-1}v)}{(v, \Lambda^{-2}v)} + \mu\left((v, v) - 1\right) + \nu\left((v, \Lambda v) - \lambda\right).$$

A necessary condition for the existence of extrema is

$$\nabla L = \frac{2}{(v, \Lambda^{-2}v)}\left(\Lambda^{-1}v - \lambda'\Lambda^{-2}v\right) + 2\mu v + 2\nu\Lambda v = 0. \tag{1.34}$$

Since $\lambda \neq \lambda_i$, $i = 1, \ldots, n$, the vector v is not collinear to any of the eigenvectors. Hence, v has at least two nonzero components v_k and v_l with $\lambda_k \neq \lambda_l$. Take k as the smallest index that

$v_k \neq 0$. Such a choice implies $\lambda_k < \lambda'$. We determine the Lagrange multipliers μ and ν from Equation (1.34) by solving the linear system

$$\begin{pmatrix} 1 & \lambda_k \\ 1 & \lambda_l \end{pmatrix} \begin{pmatrix} \mu \\ \nu \end{pmatrix} = \frac{1}{(v, \Lambda^{-2} v)} \begin{pmatrix} \lambda' \lambda_k^{-2} - \lambda_k^{-1} \\ \lambda' \lambda_l^{-2} - \lambda_l^{-1} \end{pmatrix}$$

with a nonvanishing determinant. Its unique solution reads

$$\begin{aligned} \mu &= \frac{\lambda_l^2(\lambda' - \lambda_k) + \lambda_k^2(\lambda' - \lambda_l) + \lambda_k \lambda_l \lambda'}{\lambda_k^2 \lambda_l^2 (v, \Lambda^{-2} v)}, \\ \nu &= \frac{\lambda_k \lambda_l - \lambda'(\lambda_k + \lambda_l)}{\lambda_k^2 \lambda_l^2 (v, \Lambda^{-2} v)}. \end{aligned}$$

Inserting μ and ν in the coefficient of the jth component of Equation (1.34) results in

$$\frac{1}{(v, \Lambda^{-2} v)} (\lambda_j^{-1} - \lambda' \lambda_j^{-2}) + \mu + \nu \lambda_j = -\frac{(\lambda_l - \lambda_j)(\lambda_k - \lambda_j)\rho}{\lambda_k^2 \lambda_l^2 \lambda_j^2 (v, \Lambda^{-2} v)}.$$

where

$$\rho = -\lambda_k \lambda_l \lambda_j + \lambda'(\lambda_k \lambda_l + \lambda_k \lambda_j + \lambda_l \lambda_j) > \lambda_k^2(\lambda_l + \lambda_j) > 0,$$

since $\lambda' > \lambda_k$. Hence, $v_j = 0$ for all λ_j different from λ_k and λ_l. The nonzero components v_k and v_l can be determined from $\|x\| = 1$ and $\lambda(x) = \lambda$. We obtain

$$v_k^2 = \frac{\lambda_l - \lambda}{\lambda_l - \lambda_k}, \qquad \text{and} \qquad v_l^2 = \frac{\lambda - \lambda_k}{\lambda_l - \lambda_k}. \tag{1.35}$$

Inserting these in $\lambda' = \lambda(\Lambda^{-1} v)$ results in

$$\begin{aligned} \lambda' &= \frac{\lambda_k \lambda_l (\lambda_l - \lambda_k)(\lambda_k + \lambda_l - \lambda)}{\lambda_l^3 - \lambda \lambda_l^2 + \lambda \lambda_k^2 - \lambda_k^3} \\ &= \left(\lambda_k^{-1} + \lambda_l^{-1} - (\lambda_k + \lambda_l - \lambda)^{-1} \right)^{-1} = B(\lambda_k, \lambda_l, \lambda). \end{aligned}$$

Since for $\lambda_k \leq \lambda_i < \lambda < \lambda_{i+1} \leq \lambda_l$ we have

$$\frac{\partial}{\partial \lambda_k} B(\lambda_k, \lambda_l, \lambda) > 0 \qquad \text{and} \qquad \frac{\partial}{\partial \lambda_l} B(\lambda_k, \lambda_l, \lambda) < 0,$$

so that the bound $B(\lambda_k, \lambda_l, \lambda)$ takes its maximum in $B(\lambda_i, \lambda_{i+1}, \lambda)$ while its minimum is taken in $B(\lambda_1, \lambda_n, \lambda)$ from which (1.32) follows.

The iterates of inverse iteration form a sequence of vectors with a decreasing Rayleigh quotient. It is a converging sequence since the Rayleigh quotient is bounded from below by λ_1. Hence $\lambda(x^{(k)}) - \lambda(A^{-1} x^{(k)})$ converges to 0. The residual $r(y) := Ay - \lambda(y) y$ in $y = A^{-1} x$ can be estimated from above by $\lambda(x) - \lambda(A^{-1} x)$ as follows. First, it holds

$$\begin{aligned} \|r(A^{-1} x)\|_A^2 &= \|A(A^{-1} x) - \lambda(A^{-1} x) A^{-1} x\|_A^2 \\ &= \lambda(x) - 2\lambda(A^{-1} x) + \left(\lambda(A^{-1} x) \right)^2 (x, A^{-1} x). \end{aligned} \tag{1.36}$$

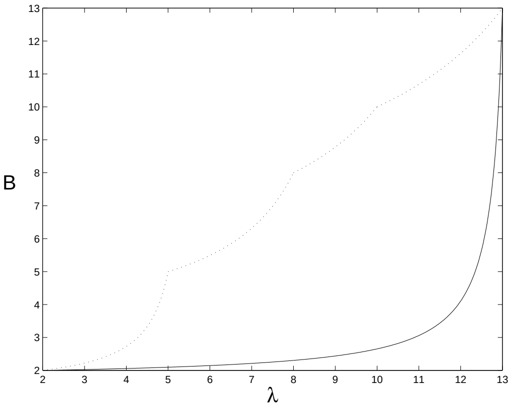


Figure 1.2:
Bounds $B(\lambda_1, \lambda_n, \lambda)$ *(solid line) and bounds* $B(\lambda_i, \lambda_{i+1}, \lambda)$*,* $i = 1, \ldots, 4$*, (dotted lines) for a model matrix with eigenvalues* $(\lambda_1, \ldots, \lambda_5) = (2, 5, 8, 10, 13)$.

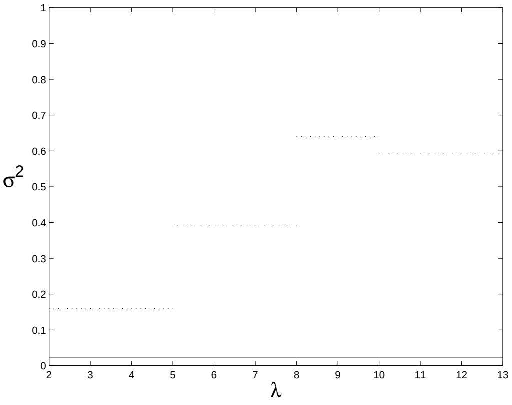


Figure 1.3:
Convergence factors σ^2 *by Corollary 1.10. Factor* $\sigma_{1,n} = \lambda_1/\lambda_n$ *(solid line) and* $\sigma_{i,i+1} = \lambda_i/\lambda_{i+1}$*,* $i = 1, \ldots, 4$*, (dotted lines) for the situation as in Figure 1.2.*

By the Cauchy-Schwarz inequality we have $(x, A^{-1}x) \leq (x, A^{-2}x)(x, x)$ and hence

$$\left(\lambda(A^{-1}x)\right)^2 (x, A^{-1}x) \leq \lambda(A^{-1}x).$$

Inserting this in (1.36) results in

$$\|r(A^{-1}x)\|_A^2 \leq \lambda - \lambda(A^{-1}x).$$

Therefore, the residuals converge to 0, i.e. the iterates converge to an eigenvector and the Rayleigh quotients converge to the corresponding eigenvalue. Finally, we note that the bound $B(\lambda_p, \lambda_q, \lambda)$ in (1.32) is attained if $x^{(k)}$ is constructed as in Equation (1.35). □

Remark 1.9. *It is well known from the linear-algebra-textbook convergence theory of inverse iteration that INVIT converges to the smallest eigenvalue* λ_1 *and to one of its corresponding eigenvectors, whenever the initial vector* $x^{(0)}$ *is not perpendicular to the invariant subspace to* λ_1*. In practice the latter condition is nearly always fulfilled due to rounding errors. Theorem 1.8 does not contain such a condition controlling the spectral components of* $x^{(0)}$*. We deliberately give up such a condition in preparation of the convergence analysis of the preconditioned eigensolvers: Any such information on spectral components of* $x^{(0)}$ *is not preserved for the PINVIT iterates (in contrast to those of INVIT!) as it is easy to construct specific preconditioners which weed out or create contributions of specific eigenfunctions in the course of the iteration. Such examples do not contradict the convergence in the sense of a decreasing Rayleigh quotient! Consequently, on the assumption on Theorem 1.8 it cannot be guaranteed that the* $\lambda^{(k)}$ *converge to the smallest eigenvalue* λ_1*. The sharp upper bound for the decrease of the Rayleigh quotients guarantees only convergence to the next smaller eigenvalue* λ_i*.*

Let us now define the ratio $\Delta_{p,q}(\lambda)$

$$\Delta_{p,q}(\lambda) := \frac{\lambda - \lambda_p}{\lambda_q - \lambda}, \tag{1.37}$$

for which Corollary 1.10 describes an upper and lower bound for the convergence factors of inverse iteration. The $\Delta_{p,q}$ factors will achieve central importance in Chapter 4 since they make a very short form of the PINVIT estimates possible.

Corollary 1.10. *It holds*

$$\Delta_{p,q}(B(\lambda_p, \lambda_q, \lambda)) = (\sigma_{p,q})^2 \, \Delta_{p,q}(\lambda). \tag{1.38}$$

Moreover, if $\lambda(x^{(k)}) \in (\lambda_i, \lambda_{i+1})$*, then*

$$\Delta_{i,i+1}(\lambda(x^{(k+1)})) \leq (\sigma_{i,i+1})^2 \Delta_{i,i+1}(\lambda(x^{(k)})) \tag{1.39}$$

and

$$\Delta_{1,n}(\lambda(x^{(k+1)})) \geq (\sigma_{1,n})^2 \Delta_{1,n}(\lambda(x^{(k)})) \tag{1.40}$$

where the convergence factor $\sigma_{p,q}$ *is given by*

$$\sigma_{p,q} = \frac{\lambda_p}{\lambda_q}.$$

Proof. To show (1.38) insert Equation (1.33) in $\Delta_{p,q}(\lambda)$. Then,

$$\begin{aligned} &\Delta_{p,q}(B(\lambda_p, \lambda_q, \lambda)) \; (\Delta_{p,q}(\lambda))^{-1} \\ &\quad = \frac{(\lambda_p^{-1} + \lambda_q^{-1} - (\lambda_p + \lambda_q - \lambda)^{-1})^{-1} - \lambda_p}{\lambda_q - (\lambda_p^{-1} + \lambda_q^{-1} - (\lambda_p + \lambda_q - \lambda)^{-1})^{-1}} \; \left(\frac{\lambda - \lambda_p}{\lambda_q - \lambda}\right)^{-1} \\ &\quad = \left(\frac{\lambda_p}{\lambda_q}\right)^2 . \end{aligned}$$

To prove (1.39) we take Equation (1.38) for $p = i$ and $q = i + 1$. Then Inequality (1.32) together with the monotonicity of $\Delta_{p,q}(\lambda)$ leads to

$$\Delta_{i,i+1}(\lambda(x^{(k+1)})) \leq \Delta_{i,i+1}(B(\lambda_i, \lambda_{i+1}, \lambda)) = (\sigma_{i,i+1})^2 \Delta_{i,i+1}(\lambda).$$

Inequality (1.40) is shown analogously. □

Remark 1.11. *The convergence factors* $\sigma_{p,q}$ *in Corollary 1.10 do not depend on* λ*. This important feature (in contrast to the bounds presented in Theorem 1.8) permits the recursive application of (1.39) so that for* $\lambda(x^{(0)}) \in (\lambda_1, \lambda_2)$ *it holds that*

$$\frac{\Delta_{1,2}(\lambda(x^{(k)}))}{\Delta_{1,2}(\lambda(x^{(0)}))} \leq \left(\frac{\lambda_1}{\lambda_2}\right)^{2k}, \qquad k = 1, 2, \ldots . \tag{1.41}$$

The chance to derive the sharp and simple recursive representation (1.41) in terms of the $\Delta_{p,q}$ ratios should be understood as the main incentive to work with these quantities in the following chapters. For instance, in Chapter 4, the representation in terms of the $\Delta_{p,q}$ allows to drastically simplify the cumbersome PINVIT(1) estimates originally given in [95, 96].

In Figure 1.3 the $\sigma_{p,q}$ factors are plotted versus $\lambda \in [\lambda_1, \lambda_n]$; in each of the intervals $[\lambda_i, \lambda_{i+1})$ they are constant. We will refer to these figures at several locations within this work since they describe the limit case of PINVIT for exact preconditioning.

Remark 1.12. *Theorem 1.8 is restricted to symmetric positive definite matrices only. To give an example of an indefinite matrix, let $A = \mathrm{diag}(-3, 1, 2)$ with $x = (1, 2, 2)^T$. Then*

$$\lambda(A^{-1}x) = 51/46 > \lambda(x) = 1,$$

as the component to $\lambda_3 = -3$ is damped out most rapidly. Therefore, inverse iteration is a descent method for the Rayleigh quotient for positive definite (semidefinite) matrices only.

2. Preconditioned inverse iteration

As a first step toward the analysis of the hierarchy of preconditioned eigensolvers, which has been introduced in Section 1.2, this chapter deals with the most simple method of this class. This is the scheme of *preconditioned inverse iteration*, which derives from Algorithm 1.5 for $k = s = 1$ and whose convergence theory has been given in [95, 96]. The aim of this chapter is to review the central ideas underlying its convergence analysis: In Section 2.1.1 we first compile the assumptions made on the preconditioners as used throughout this work. These assumptions are typically fulfilled for (scaled) preconditioners based on classical multigrid or domain decomposition schemes and therefore do not confine the generality of the approach presented here. This is followed in Section 2.1.2 by the derivation of some convenient normal form of these preconditioners, which turns out to be very useful, both for the representation of our eigensolvers as well as for their analysis.

The key idea of the convergence analysis is to disregard the somewhat unpleasant behavior of *single* preconditioners, but to apply (the set of) *all* admissible preconditioners to the actual iterate, which results in a corresponding set of new iterates. The latter set, as a consequence of the assumptions on the preconditioners, turns out as a ball (with respect to the A-norm) whose center is given by the result of inverse iteration if applied to the actual iterate, see Section 2.2.

A detailed insight into this geometry makes possible the localization of extremum points of the Rayleigh quotient on these balls and proves as a valuable tool for the convergence analysis, compare Section 2.4. In this chapter we restrict the analysis to vector schemes; corresponding subspace schemes are described in Chapter 5 and [98].

2.1 Preconditioning for eigenvalue solvers

Preconditioning techniques for the solution of large *linear systems* of equations, as arising from the discretization of (elliptic) partial differential operators in mathematical physics, are well accepted tools to guarantee rapid convergence. Contrastingly, comparable preconditioned *eigensolvers* are relatively little known, though quite successful schemes have been developed, e.g. eigensolvers based on multigrid preconditioning [75] with successful applications to $\mathcal{O}(N)$ density-functional theory methods [40], to structural mechanics [21] and to the Maxwell equations [60], to mention only some areas. For a more detailed discussion, see

Chapter 1, where further examples are reviewed showing that preconditioning in eigenvalue computations is not limited to multigrid preconditioners.

Throughout this work we do not consider or construct special preconditioners for the eigenproblem but use preconditioners originally designed for the solution of linear systems. Our aim is to show that these preconditioners can successfully be applied to solve eigenvalue problems resulting in highly efficient, stable and robust algorithms that converge with a grid-independent rate.

Let us now point out some differences between optimal preconditioning (in the sense of fastest convergence) for systems of *linear equations* and optimal preconditioning for *eigenproblems*:

- Optimal preconditioning for the solution of linear systems $Ax = b$ is done by the inverse matrix $B^{-1} = A^{-1}$, which results in one-step convergence to the *exact* solution. In the following we refer to this choice as *exact preconditioning*.

- In contrast to this, the choice $B^{-1} = A^{-1}$ as a preconditioner for eigenvalue problems is obviously not optimal, since this will reduce Algorithm 1.5 to Subspace Iteration as given by Algorithm 1.4. It is important to note that we are not interested in an exact solution of the system of linear equations associated with inverse iteration, but that we want to compute the best possible eigenvalue/vector approximation that can be achieved by the admissible preconditioners. The best preconditioner for the partial eigenvalue problem (to determine the smallest eigenvalue λ_1 together with the eigenvector x_1), would result in a new iterate collinear to x_1. Here we treat the question of how "close" such an optimal preconditioner for the eigenvalue problem is to the set of admissible preconditioners.

  Surely, the action of any (algebraic) eigensolver, which computes an approximation to x_1, can be associated with an approximation to such an ideal preconditioner. But here we do not deal with such ideal but usually expensive and in some sense unrealistic preconditioning. In our setup *optimal preconditioning* is done by that (unique) preconditioner in the set of admissible preconditioners, which is responsible for the fastest decrease of the Rayleigh quotient of the new iterate. A detailed analysis of this optimal preconditioning is given in Chapter 3.

As a notable feature and strength of the present analysis, it completely separates the questions of the choice of the preconditioner and that of the linear algebra of the eigensolvers. Hence, there is no need to discuss the construction or the underlying principles of preconditioners. Let us briefly mention some important classes of preconditioners, e.g. preconditioners based on classical (algebraic) multigrid methods or on domain decomposition schemes as well as those based on incomplete factorizations. Several of these preconditioners have reached practical importance in scientific and industrial applications. We refer to Bramble [14] for multigrid preconditioning and to Saad [118] for a detailed introduction to preconditioning techniques for general linear systems.

Here, we do not consider shift-and-invert preconditioning [117], i.e. we only use approximate inverses of the system matrix A and not those of a shifted matrix $A - \sigma I$. Preconditioned eigensolvers based on shift-and-invert techniques are characterized by a relatively fast convergent outer loop iteration (for instance the shifted inverse iteration, the Rayleigh quotient iteration or even the Arnoldi procedure). But the major drawback of these methods is to be seen in the fact, that in the inner loop, they require a high accuracy solution of linear systems in nearly singular matrices. As a central ingredient of such schemes one has to devise a proper stopping condition for the inner iteration. Finding and implementing such conditions is a nontrivial task. Their appropriate choice appears decisive for the effectiveness of the inner-outer loop scheme. These drawbacks hamper the efficiency of the eigensolver, lead to an increased algorithmic complexity and result in a loss of practical robustness.

2.1.1 Assumptions on the preconditioners

The preconditioner (or approximate inverse) B^{-1} for A is assumed to be a symmetric positive definite matrix, which approximates the inverse of A in such a way that

$$(1-\gamma)(x, Bx) \leq (x, Ax) \leq (1+\gamma)(x, Bx), \qquad \text{for all } x \in \mathbb{R}^n, \tag{2.1}$$

where γ is a positive constant less than 1. We can rewrite (2.1) in two alternative, but equivalent forms: On the one hand, (2.1) says that

$$\kappa_2(B^{-1}A) \leq \frac{1+\gamma}{1-\gamma}$$

for the spectral condition number κ_2 of the preconditioned matrix $B^{-1}A$. On the other hand, (2.1) provides a bound for the operator norm induced by A (or, alternatively, for the spectral radius) of the error propagation matrix $I - B^{-1}A$ in the form

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

Spectral assumptions like (2.1) are typical for multigrid or domain decomposition preconditioners, leaving aside the fact that γ is not always amenable in practice, e.g. if one wants to apply preconditioners based on incomplete (Cholesky) factorization. Nevertheless, even in the latter cases the spectral condition number $\kappa_2(B^{-1}A)$ is a measure for the quality of the preconditioner.

In the case of mesh eigenvalue problems we assume γ to be *independent of the mesh size,* which holds for the best multigrid/domain decomposition preconditioners [14]. Weakening the latter assumption, i.e. allowing a slight dependence of γ on the mesh size would lead to a comparable dependence for the convergence estimate of the associated preconditioned eigensolver.

Throughout this work we make use of the following definition of the set $\mathcal{B}_\gamma$ containing all admissible preconditioners, i.e. those satisfying the condition (2.2),

$$\mathcal{B}_\gamma = \{B^{-1} \in \mathbb{R}^{n\times n} : \ B \text{ symmetric positive definite, } ||I - B^{-1}A||_A \leq \gamma\}. \tag{2.3}$$

As we will see, the set $\mathcal{B}_\gamma$ reflects the simple geometry underlying PINVIT, cf. the definition by (2.12). The assumption (2.1) has not only been used in the convergence analysis of PINVIT in [95, 96], but also for the analysis of the corresponding subspace scheme in [15, 98].

Preconditioners are sometimes characterized by the somewhat more general spectral equivalence of the form

$$\delta_0(x, Bx) \leq (x, Ax) \leq \delta_1(x, Bx), \qquad \text{for all } x \in \mathbb{R}^n, \tag{2.4}$$

for positive constants δ_0 and δ_1.

As long as $\delta_1 < 2$ one can reformulate (2.4) in such a way that (2.1) holds, but this would result in a loss of sharpness if $\delta_0 \neq 2 - \delta_1$. In general, the preconditioner can be scaled by

$$\vartheta = \frac{2}{\delta_0 + \delta_1}, \tag{2.5}$$

which leads to the smallest possible bound

$$||I - \vartheta B^{-1}A||_A \leq \frac{\delta_1 - \delta_0}{\delta_0 + \delta_1} < 1. \tag{2.6}$$

In most situations ϑ is not explicitly available. But note that the knowledge of ϑ is only required for the most simple scheme PINVIT(1), since for the improved techniques PINVIT(k,s) with $k \geq 2$ the choice of $\vartheta \neq 0$ is immaterial. This independence holds trivially, since in Algorithm 1.5 any (nonzero) scaling of $\tilde{U}_{j+1}$ has no influence on the result of the Rayleigh-Ritz procedure. We summarize these considerations in Lemma 2.1.

Lemma 2.1. *Scaling of preconditioners for the schemes PINVIT(k,s) for $k \geq 2$ is immaterial. Hence, for $k \geq 2$ the convenient assumption (2.2) means no loss of generality.*

In applications with a preconditioner satisfying only (2.4) (or whenever one is unsure about the validity of (2.1)), we recommend to use the improved schemes PINVIT(k,s) for $k = 2, 3$. For slightly higher costs (compared to $k = 1$) one can remedy the shortcoming of (2.4) and is rewarded with improved convergence properties.

2.1.2 Normal form of preconditioners

As a preparatory step for the following analysis, we introduce in Lemma 2.2 some normal form of the preconditioners contained in $\mathcal{B}_\gamma$. Lemma 2.2, which generalizes Lemma 2.2 in [95], shows that the degrees of freedom for constructing B^{-1} consist in choosing some orthogonal matrix V and some diagonal matrix D with bounded diagonal elements.

Lemma 2.2. *Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric positive definite matrices with*

$$\|I - B^{-1}A\|_A \leq \gamma \tag{2.7}$$

for $0 \leq \gamma < 1$. Then there is an orthogonal matrix $V \in \mathbb{R}^{n\times n}$ and a diagonal matrix $D = \mathrm{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n\times n}$ with $|d_i| \leq 1$, $i = 1, \ldots, n$, so that

$$B^{-1} = A^{-1} + \gamma A^{-1/2} V D V^T A^{-1/2}. \tag{2.8}$$

If $\max_k |d_k| = 1$, then $\|I - B^{-1}A\|_A = \gamma$.

Proof. Let $B^{-1} = A^{-1} + Z$. Then we have

$$\|I - B^{-1}A\|_A = \|A^{1/2} Z A^{1/2}\|_2$$

where $A^{1/2}ZA^{1/2}$ is a symmetric matrix. Hence, $A^{1/2}ZA^{1/2} = VDV^T$ for some orthogonal V and diagonal D. The rest of the lemma can easily be verified. □

In Equation (2.8) the preconditioner B^{-1} is generated by "perturbing" the inverse of A by $\gamma A^{-1/2}VDV^TA^{-1/2}$. The latter term is generated by VDV^T for some orthogonal matrix $V \in \mathbb{R}^{n\times n}$ and an appropriate diagonal matrix D. If we substitute VDV^T by some low-rank modification of the identity matrix having the form $I - 2VDV^T$ for orthonormal $V \in \mathbb{R}^{n\times k}$, $1 \leq k \leq n$, then we obtain an alternative preconditioner

$$B^{-1} := A^{-1} + \gamma A^{-1/2}(I - 2VDV^T)A^{-1/2}. \tag{2.9}$$

Setting $D = \mathrm{diag}(d_1, \ldots, d_k)$ with $|d_i| \leq 1$ results in a symmetric positive definite preconditioner which satisfies (2.2). These preconditioners will be useful in many respects in the following analysis. As it will turn out later, the suprema of the Rayleigh quotient of the PINVIT(k) schemes are located on the surface of the set of possible iterates. These extrema can be generated by preconditioners for which D, in (2.9), equals the identity matrix $I \in \mathbb{R}^{k\times k}$. Note that the choice $D = I$ in (2.8) for orthonormal $V \in \mathbb{R}^{n\times k}$, $k < n$, is not appropriate to generate the full surface, as VV^T is a projector, while $I - 2VV^T$ in (2.9) is a reflection.

Reducing the degrees of freedom of (2.9) to a minimum results in preconditioners built from Householder reflections $H = I - 2vv^T$, $\|v\| = 1$, in a way that

$$B^{-1} = A^{-1} + \tilde{\gamma} A^{-1/2} H A^{-1/2} \tag{2.10}$$

for $0 \leq \tilde{\gamma} \leq \gamma$. Obviously, the choice $\tilde{\gamma} = \gamma$ causes a bijection of the surface of the unit ball, i.e. $\|v\| = 1$, to the surface of the set of possible PINVIT iterates. (Just note that we will exploit a comparable relation for PINVIT(2) later in Section 6.5.3.)

For a formal description of the set of iterates that can be attained by the PINVIT(k) schemes, apply *all* preconditioners in $\mathcal{B}_\gamma$ to a fixed x, i.e. consider the mapping

$$\mathcal{B}_\gamma \to E_\gamma^k(x) : B^{-1} \mapsto PINVIT(k)[x]. \tag{2.11}$$

Therein $PINVIT(k)[x]$ denotes the new iterate according to Algorithm 1.5, and $E_\gamma^k(x)$ is defined to be the set of possible iterates. In this chapter we review some properties of the set $E_\gamma^1(x)$ and the implications for the convergence of the associated eigensolver. Later in Chapter 5 we will present an analysis of PINVIT(2) for which a detailed description of E_γ^2 is given. For the sake of simplicity we abbreviate $E_\gamma^1(x)$ by $E_\gamma(x)$ (or even shorter E_γ) and we write explicitly

$$E_\gamma(x) := \{x - B^{-1}(Ax - \lambda x) : \ B^{-1} \in \mathcal{B}_\gamma\}, \tag{2.12}$$

as well as the corresponding iterative scheme

$$x' = x - B^{-1}(Ax - \lambda x) \tag{2.13}$$

of preconditioned inverse iteration as derived from Algorithm 1.5 for $k = s = 1$.

As a first observation, note that $E_\gamma(x)$ is a closed convex ball whose center is given by the result of (scaled) inverse iteration.

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

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

Proof. By definition (2.12) the set $E_\gamma(x)$ is a subset of the ball. To show the opposite inclusion consider a point $\lambda A^{-1}x + y$ in the ball. Let $\tilde{\gamma} := (\|y\|_A/\|(I - \lambda A^{-1})x\|_A)$ so that $\tilde{\gamma} \leq \gamma$. Then a Householder reflection H can be determined in a way that

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

Inserting this H in (2.10) and applying that preconditioner to (2.13) results in $\lambda A^{-1}x + y$ since

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

□

Lemma 2.3 gives rise to the idea to substitute the somewhat intricate analysis of scheme (2.13) involving the preconditioners in $\mathcal{B}_\gamma$ by the *much simpler problem to identify points of poorest convergence in the ball $E_\gamma(x)$*. We use the Rayleigh quotient as the convergence measure and treat the problem to locate its (unique) point of a supremum in $E_\gamma(x)$. This point is associated with a preconditioner of poorest convergence—but as the decisive advantage of this approach, there is no necessity to consider the explicit form of that preconditioner responsible for the poorest convergence.

2.2 A geometric representation

Trying to reformulate the assumptions on the set of admissible preconditioners as geometric conditions for the set of iterates $E_\gamma(x)$, we summarize in Lemma 2.4 some immediate consequences of the fact that $E_\gamma(x)$ is a ball with respect to the A-geometry. First of all a fundamental A-orthogonal decomposition is shown, which implies that the origin is not contained in $E_\gamma(x)$. This fact is a necessary prerequisite to guarantee convergence since in any neighborhood of 0 (not including the origin itself) the Rayleigh quotient takes its full range $[\lambda_1, \lambda_n]$.

Lemma 2.4. *For $x \in \mathbb{R}^n \setminus \{0\}$ it holds that*

$$0 = (x, (I - \lambda A^{-1})x)_A, \tag{2.14}$$

$$\|\lambda A^{-1}x\|_A^2 = \|x\|_A^2 + \|(I - \lambda A^{-1})x\|_A^2, \tag{2.15}$$

$$0 \notin E_\gamma(x) \quad \textit{for all } \gamma \in [0, 1]. \tag{2.16}$$

Proof. Equations (2.14) and (2.15) 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 (2.15) for nonzero x result in

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

□

Figure 2.1 illustrates the set $E_\gamma(x)$ within the plane $\operatorname{span}\{x, \lambda A^{-1}x\}$ as well as the A-orthogonal decomposition (2.15).

Our next aim is to simplify the representation of the scheme (2.13). Therefore we adopt the common practice to carry out the analysis within a basis of eigenvectors of A, which essentially means that A is assumed to be a diagonal matrix. Obviously, it cannot be assumed that this basis diagonalizes the preconditioner, too. Here, we apply a slightly different transformation, i.e. we transform PINVIT to the *A-orthonormal basis of eigenvectors of A*, i.e. we scale the Euclidean-orthonormal eigenvectors of A by the factors $1/\lambda_i^{1/2}$, $i = 1, \ldots, n$. We call this new basis the *c-basis* and the initial basis the *x-basis*.

Definition 2.5 (The c-basis representation). *Let X be the orthogonal matrix containing the eigenvectors of A in its columns so that $X^TAX = \Lambda$ and $X^TX = I$. The diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ contains the eigenvalues of A. We define the coefficient vector c of x with respect to the basis of A-orthonormal eigenvectors of A by*

$$x = X\Lambda^{-1/2}c. \tag{2.17}$$

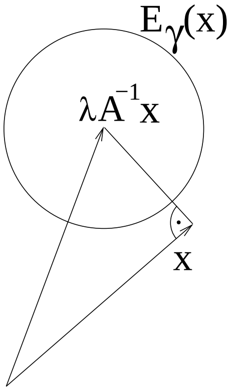


Figure 2.1: $E_\gamma(x)$ *sliced along* $\mathrm{span}\{x, \lambda A^{-1}x\}$ *w.r.t.* $\|\cdot\|_A$*-geometry.*

We immediately obtain for the Rayleigh quotient of some coefficient vector $d \in \mathbb{R}^n$ with respect to the c-basis the following representation

$$\lambda(d) = \frac{(d,d)}{(d,\Lambda^{-1}d)}, \tag{2.18}$$

which is checked by writing out the Rayleigh quotient (1.3) of $X\Lambda^{-1/2}d$. We usually abbreviate $\lambda := \lambda(c)$.

We highlight the following properties of the c-basis representation:

1. The c-basis representation of the discretization matrix equals the identity matrix because of
$$(X\Lambda^{-1/2})^T AX\Lambda^{-1/2} = I.$$
In the same way the identity is transformed to Λ^{-1}. As a pleasant feature, E_γ turns out as a ball with respect to the Euclidean geometry.

2. While the gradient of the Rayleigh quotient $\lambda(x) = (x,Ax)/(x,x)$ within the x-basis, as given by
$$\nabla\lambda(x) = \frac{2}{(x,x)}(Ax - \lambda x)$$
is not directed to the center of $E_\gamma(x)$ as $\nabla\lambda(x)$ is not collinear to $x - \lambda A^{-1}x$, the gradient vector of (2.18)
$$\nabla\lambda(c) = \frac{2}{(c,\Lambda^{-1}c)}(c - \lambda\Lambda^{-1}c),$$
points to the center $\lambda\Lambda^{-1}c$ of the ball. This property sets up an appropriate geometry and turns out as decisive for several respects in the analysis of (2.13).

Let us now reformulate PINVIT within the c-basis based on preconditioners of the form (2.10) as they span the full ball E_γ.

Lemma 2.6. *Preconditioned inverse iteration for the preconditioner (2.10) takes (with respect to the A-orthonormal basis of eigenvectors of A) the form*

$$c' = \lambda\Lambda^{-1}c - \tilde{\gamma}(I - 2vv^T)(I - \lambda\Lambda^{-1})c, \tag{2.19}$$

where c and c' are the coefficient vectors within this basis of x and x', respectively. All admissible preconditioners are spanned for $0 \leq \tilde{\gamma} \leq \gamma$ and $v \in \mathbb{R}^n$, $\|v\| = 1$.

Proof. Inserting (2.17) in (2.13) and using (2.10) we obtain

$$c' = c - \Lambda^{1/2}X^TB^{-1}X\Lambda^{1/2}(I - \lambda\Lambda^{-1})c = \lambda\Lambda^{-1}c - \tilde{\gamma}X^THX(I - \lambda\Lambda^{-1})c, \tag{2.20}$$

so that (2.19) follows since both H and X^THX are Householder reflections. □

For the sake of convenience we define $E_\gamma(c)$ to be the c-basis representation of $E_\gamma(x)$, i.e.

$$E_\gamma(c) := \{\Lambda^{1/2}X^Tz : \; z \in E_\gamma(x)\} = \{c' \text{ given by (2.19)}\}. \tag{2.21}$$

We conclude this section with the important remark that the maximal Rayleigh quotient on $E_\gamma(c)$ does not depend on the signs of the components of c, because a change of the sign of the kth component of c corresponds to a reflection of $E_\gamma(c)$ by a hyperplane orthogonal to the kth unit vector through the origin. Since the Rayleigh quotient (2.18) is a purely quadratic function in the components of its argument, any sign dependence vanishes and the Rayleigh quotient takes the same values on the reflected ball.

This gives us the justification to restrict the convergence analysis to non-negative coefficient vectors c.

2.3 Multiple eigenvalues

The aim of this section is to provide a justification for restricting the convergence analysis of preconditioned inverse iteration to matrices having only simple eigenvalues. In this section, we therefore assume A to be a real symmetric $m \times m$ matrix with n different eigenvalues $0 < \lambda_1 < \ldots < \lambda_n$. The multiplicity of λ_i is given by m_i so that $m = \sum_{i=1}^n m_i$. Then within the c-basis the diagonal matrix Λ reads

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

We write the corresponding coefficient vectors as

$$d = (d_{1,1}, \ldots, d_{1,m(1)}, \ldots \quad , d_{n,1}, \ldots, d_{n,m(n)})^T \in \mathbb{R}^m,$$

where $d_{i,j}$ denotes the jth component corresponding to the ith eigenvalue of multiplicity $m(i)$. Now consider the mapping $P : \mathbb{R}^m \to \mathbb{R}^n$, which defines a corresponding eigenvalue problem of a smaller dimension with the same but simple eigenvalues by condensing components belonging to a multiple eigenvalue in a joint component.

$$(Pd)|_i = \bar{d}_i := (\sum_{j=1}^{m(i)} d_{i,j}^2)^{1/2}. \tag{2.22}$$

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

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

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

$$\bar{c}' = \bar{\lambda}(\bar{c})\bar{\Lambda}^{-1}\bar{c} - \tilde{\gamma}\bar{H}(I - \bar{\lambda}(\bar{c})\bar{\Lambda}^{-1})\bar{c} \tag{2.23}$$

for arbitrary Householder reflections $\bar{H} \in \mathbb{R}^{n \times n}$. Obviously, (2.23) defines a ball $E_\gamma(\bar{c}) \subset \mathbb{R}^n$. The next lemma shows that the suprema in the case of simple eigenvalues dominate those of the multiple eigenvalue case.

Lemma 2.7. *Let $c \in \mathbb{R}^m$, then*

$$\sup \lambda(E_\gamma(c)) \leq \sup \bar{\lambda}(E_\gamma(\bar{c})). \tag{2.24}$$

Proof. First observe that P keeps the Rayleigh quotient invariant in a sense that

$$\lambda(d) = \bar{\lambda}(Pd), \qquad d \in \mathbb{R}^m. \tag{2.25}$$

Especially, $\lambda = \lambda(c) = \bar{\lambda}(Pc)$ so that P maps the center of $E_\gamma(c)$ to the center of $E_\gamma(\bar{c})$, i.e. $P(\lambda\Lambda^{-1}c) = \lambda\bar{\Lambda}^{-1}\bar{c}$. Because of $\|c - \lambda\Lambda^{-1}c\| = \left\|\bar{c} - \lambda\bar{\Lambda}^{-1}\bar{c}\right\|$ both balls have the same radius.

Since for any $d, e \in \mathbb{R}^m$ (with $\bar{d} = Pd$ and $\bar{e} = Pe$) we have (by using the Cauchy-Schwarz inequality)

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

from which we conclude that $P(E_\gamma(c))$ is a subset of $E_\gamma(\bar{c})$. Therefore, it holds

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

and the proposition follows with (2.25). □

Inequality (2.24) is sharp as shown in Section 3.3 of [96]. We conclude that we do not lose generality by restricting the analysis of preconditioned inverse iteration to matrices having only simple eigenvalues.

2.4 A convergence theorem

Having reviewed the geometry underlying the analysis of preconditioned inverse iteration in the previous sections, let us now reproduce the central convergence theorem from [95, 96], where the reader will also find the complete convergence proof. We note that analogous proof techniques are employed for the analysis of the fastest possible convergence in Chapter 3.

Let us summarize the main steps of the convergence proof:

1. Poorest preconditioning: For a given vector c find that point w in $E_\gamma(c)$ in which the Rayleigh quotient takes its maximum. Then $w = w[c]$ is considered as the point of poorest convergence with respect to the choice of the preconditioner (from the set $\mathcal{B}_\gamma$ of admissible preconditioners). Then for the Rayleigh quotient $\lambda(w)$ it holds that

   $$\lambda(w) < \lambda(c)$$

   and the decrease $\lambda(c) - \lambda(w)$ is considered as a convergence measure toward the next smaller eigenvalue.

2. Level set dependence: There is another degree of freedom to be eliminated. Having given an iteration vector with the Rayleigh quotient $\lambda \in (\lambda_1, \lambda_n)$, its expansion in eigenfunctions is not unique. In other words for a fixed λ the Rayleigh quotient $\lambda(w)$ depends on the choice of c from the level set

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

   Hence, the second task is to determine the set

   $$\arg \max_{c \in L(\lambda)} \lambda(w[c]) \tag{2.26}$$

   of vectors of poorest convergence with respect to the level set $L(\lambda)$. As shown in [96] the set (2.26) is spanned by a unique vector aside from scaling and the signs of the contributing eigenvectors.

3. Mini-dimensional analysis: In a final step we derive the Rayleigh quotient $\lambda(w[c^*])$ for a vector c^* of poorest convergence where c^* is given by (2.26). As c^* is contained in a 2D invariant subspace of A (compare to a similar property of inverse iteration), we determine $\lambda(w[c^*])$ by some mini-dimensional analysis. The latter value serves as a sharp upper bound for $\lambda(c')$ by (2.19) for an arbitrary choice of $c \in L(\lambda)$ *and* for all admissible preconditioners.

Here we do not discuss convergence estimates for the residual. Those bounds can be derived for instance by the Temple inequality and they show that the residual converges to 0 if the Rayleigh quotient tends to the eigenvalue λ_1. We refer to Section 3.2 in [96] concerning convergence estimates for the eigenvector approximations. The main reason for our reticence is that the sequence of the acute angles between the eigenvector belonging to λ_1 and the PINVIT iterates is not always monotone decreasing.

In Theorem 2.8 we summarize the results of the convergence analysis of (2.13) in terms of a sharp estimate for the stepwise decrease of the Rayleigh quotient. In other words, we give an upper estimate for the Rayleigh quotient $\lambda' = \lambda(x')$ of the new iterate. This bound depends not only on the Rayleigh quotient $\lambda = \lambda(x)$ of the actual iterate but also on γ, λ_i and λ_{i+1} if $\lambda \in (\lambda_i, \lambda_{i+1})$. As a measure for the relative decrease of λ' toward the next smaller eigenvalue λ_i, we consider the ratio

$$\Phi_{i,i+1}(\lambda,\gamma) := \frac{\lambda' - \lambda_i}{\lambda - \lambda_i}.$$

In the case of poorest convergence it always holds $\lambda' > \lambda_i$, which guarantees positiveness of $\Phi_{i,i+1}(\lambda,\gamma)$. The following theorem will show that this ratio is smaller than 1. To make the theorem easily accessible to the reader, it is formulated with respect to the initial basis (the x-basis) and not with respect to the more technical c-basis.

Theorem 2.8. *Let $x^{(0)} \neq 0$ be an initial vector with the Rayleigh quotient $\lambda^{(0)} := \lambda(x^{(0)})$ and denote the sequence of iterates of preconditioned inverse iteration (2.13) by*

$$(x^{(k)}, \lambda^{(k)}), \qquad k = 0, 1, 2, \ldots,$$

where $\lambda^{(k)} = \lambda(x^{(k)})$. The preconditioner is assumed to satisfy (2.2) for some $\gamma \in [0, 1)$.

Then the sequence of Rayleigh quotients $\lambda^{(k)}$ decreases monotonically and $(x^{(k)}, \lambda^{(k)})$ converges to an eigenpair of A. Moreover, let $(x, \lambda) = (x^{(k)}, \lambda(x^{(k)}))$ be the iterates of the kth step, $k \geq 0$, and denote the new iterates by $(x', \lambda') = (x^{(k+1)}, \lambda(x^{(k+1)}))$. Then it holds that:

1. *For $\lambda = \lambda_1$ or $\lambda = \lambda_n$ the iteration is stationary in an eigenvector of A.*
   If $\lambda = \lambda_i$, with $2 \leq i \leq n-1$, then λ' takes its maximal value $\lambda' = \lambda_i$ if (2.13) is applied to the eigenvector x_i belonging to the eigenvalue λ_i.

2. *If $\lambda_i < \lambda < \lambda_{i+1}$, then λ' takes its maximum with respect to $x \in L(\lambda)$ in the vector $x = x_{i,i+1}$ with*

   $$x_{i,i+1} = \omega_1 x_i + \omega_2 x_{i+1},$$

   for suitable real constants ω_1 and ω_2. In other words, $x_{i,i+1}$ is contained in the invariant subspace spanned by the eigenvectors x_i and x_{i+1}. The supremum concerning poorest *preconditioning $B^{-1} \in \mathcal{B}_\gamma$ leads to the Rayleigh quotient $\lambda' = \lambda_{i,i+1}(\lambda,\gamma)$ where*

$$\begin{aligned}\lambda_{i,j}(\lambda,\gamma) \;=\;& \lambda\lambda_i\lambda_j(\lambda_i+\lambda_j-\lambda)^2 / \\ & \Big(\gamma^2(\lambda_j-\lambda)(\lambda-\lambda_i)(\lambda\lambda_j+\lambda\lambda_i-\lambda_i^2-\lambda_j^2) \\ & -2\gamma\sqrt{\lambda_i\lambda_j}(\lambda-\lambda_i)(\lambda_j-\lambda) \\ & \quad \sqrt{\lambda_i\lambda_j+(1-\gamma^2)(\lambda-\lambda_i)(\lambda_j-\lambda)} \\ & -\lambda(\lambda_i+\lambda_j-\lambda)(\lambda\lambda_j+\lambda\lambda_i-\lambda_i^2-\lambda_i\lambda_j-\lambda_j^2)\Big)\,.\end{aligned} \tag{2.27}$$

For the relative decrease of $\lambda' = \lambda_{i,i+1}(\lambda,\gamma)$ toward the nearest eigenvalue λ_i smaller than λ it holds

$$\Phi_{i,i+1}(\lambda,\gamma) = \frac{\lambda_{i,i+1}(\lambda,\gamma)-\lambda_i}{\lambda-\lambda_i} < 1. \tag{2.28}$$

The proof of Theorem 2.8 is given in [95, 96].

The reader should not be confused by the complex form of the bound (2.27) and should understand its complexity as the price one has to pay for having an estimate that is sharp in λ, λ_i, λ_{i+1} and γ. The major drawback of (2.27) is that the dependence on its arguments is not “visibly” clear. A remedy overcoming this disadvantage is given in Chapter 4 where a simple and short convergence estimate is derived by eliminating the dependence on λ, which essentially means that we sacrifice the sharpness in λ.

Theorem 2.8 guarantees that the Rayleigh quotients of the iterates form a *monotone decreasing sequence of numbers converging to an eigenvalue*: To elucidate the convergence behavior in a more detailed way, consider the sequence of iterates in the form $(x^{(k)}, \lambda^{(k)})$, for $k = 0, 1, 2, \ldots$. If one starts with an initial eigenvalue approximation larger than (the possibly interior eigenvalue) λ_i, it cannot be said in principle when the Rayleigh quotients $\lambda^{(k)}$ move from one interval $[\lambda_i, \lambda_{i+1})$ to the next interval of smaller eigenvalues and finally to the “catchment basin” $[\lambda_1, \lambda_2)$ of the smallest eigenvalue λ_1. For the moment we assume the Rayleigh quotients to have reached the interval $[\lambda_1, \lambda_2)$. Then the “one-step” estimates $\Phi_{i,i+1}$ for $i = 1$ can be used to define a *convergence rate* estimate $\Theta_{1,2}(\lambda,\gamma)$ for PINVIT

$$\Theta_{1,2}(\lambda,\gamma) := \sup_{\lambda_1<\tilde{\lambda}\le\lambda} \Phi_{1,2}(\tilde{\lambda},\gamma), \qquad \lambda \in (\lambda_1, \lambda_2]. \tag{2.29}$$

(One should note that $\Theta_{1,2}(\lambda,\gamma)$ only slightly differs from $\Phi_{1,2}(\lambda,\gamma)$. In the example discussed below, see Figure 2.2, the curve $\Phi_{1,2}(\lambda, 0.9)$ in the interval $[2,5]$ takes its minimum in $\lambda \approx 2.44$ instead of $\lambda = 2$.)

Now, $\Theta_{1,2}(\lambda,\gamma)$ is an upper bound for the relative decrease of the Rayleigh quotients in the sense that

$$\frac{\lambda^{(k+1)}-\lambda_1}{\lambda^{(k)}-\lambda_1} \le \Theta_{1,2}(\lambda,\gamma), \qquad \text{for } k = 1, 2, \ldots.$$

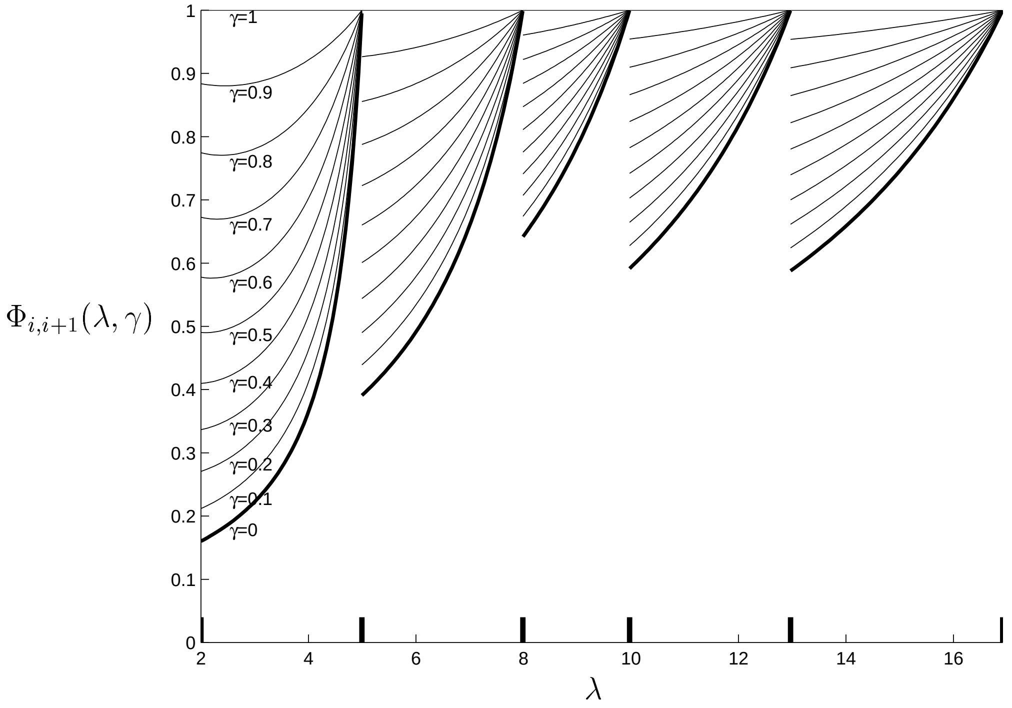


Figure 2.2: *Convergence estimates* $\Phi_{i,i+1}(\lambda,\gamma)$ *for the 10 smallest eigenvalues* $\lambda_h^{(k,l)}$ *given by Equation (2.30).*

In other words, the Rayleigh quotients $\lambda^{(k)}$ converge *linearly* to λ_1 with the convergence rate $\Theta_{1,2}$.

We lay special emphasis on the fact that $\Theta_{1,2}(\lambda,\gamma)$ for a mesh eigenproblem can be *bounded away from 1 independently of the mesh size,* see the discussion on grid-independent convergence at the end of this chapter.

Let us now explain and illustrate the results by discussing the five-point finite difference discretization of the eigenproblem for the Laplacian on the square $[0,\pi]^2$ with homogeneous Dirichlet boundary conditions. The eigenvalues of the continuous problem $\lambda^{(k,l)}$ and of the finite difference discretization $\lambda_h^{(k,l)}$, for the mesh size h, $h=\pi/N$ for $N\in\mathbb{N}$, are given by

$$\lambda^{(k,l)} = k^2{+}l^2, \qquad \lambda_h^{(k,l)} = \frac{4}{h^2}\left(\sin^2(\frac{kh}{2}) + \sin^2(\frac{lh}{2})\right), \qquad k,l=1,2,\ldots,N{-}1. \tag{2.30}$$

The 10 smallest eigenvalues $\lambda^{(k,l)}$ including multiplicity read $2,5,5,8,10,10,13,13,17,17$. For $h=\pi/50$ these eigenvalues coincide with the $\lambda_h^{(k,l)}$ within the 1 percent range. Figure 2.2 shows the convergence estimates $\Phi_{i,i+1}(\lambda,\gamma)$ plotted against $\lambda\in[2,17]$ for eleven values of γ, $\gamma=0,0.1,\ldots,1$. The eigenvalues $\lambda_i := \lambda_h^{(k,l)}$ are marked by bold vertical lines on the abscissa. Note that by Section 2.3 there is no dependence on the multiplicity of the eigenvalues.

The bold curves represent the case $\gamma = 0$, i.e. $B = A$, for which PINVIT is identical with the inverse iteration procedure (INVIT). We explicitly derive the estimate of poorest INVIT convergence, by inserting $\gamma = 0$ and $j = i + 1$ in (2.27) and obtain

$$\lambda_{i,i+1}(\lambda, 0) = \left(\lambda_i^{-1} + \lambda_{i+1}^{-1} - (\lambda_i + \lambda_{i+1} - \lambda)^{-1}\right)^{-1}, \tag{2.31}$$

and

$$\Phi_{i,i+1}(\lambda, 0) = \frac{\lambda_i^2}{\lambda_i^2 + (\lambda_{i+1} - \lambda)(\lambda_i + \lambda_{i+1})}. \tag{2.32}$$

In each interval $[\lambda_i, \lambda_{i+1})$ inverse iteration (as shown in Theorem 1.8) attains its poorest convergence in those vectors which are spanned by the eigenvectors corresponding to λ_i and λ_{i+1}.

For $\lambda = \lambda_{i+1}$ we have $\Phi_{i,i+1}(\lambda_{i+1}, \gamma) = 1$, which expresses the fact that INVIT and PINVIT are stationary in the eigenvectors of A. The curves in Figure 2.2 for $\gamma > 0$ describe the case of poorest convergence of PINVIT. For $\gamma = 1$ poorest convergence means stationarity of the eigensolver, i.e.

$$\Phi_{i,i+1}(\lambda, 1) = 1 \quad \text{for } \lambda \in [\lambda_i, \lambda_{i+1}] \text{ and } i = 1, \dots, n-1.$$

For decreasing γ PINVIT behaves more and more like inverse iteration. As a result of Theorem 2.8 poorest convergence is attained in $x_{i,i+1}$, which underlines the close relation of inverse iteration and preconditioned inverse iteration.

In principle, it cannot be guaranteed that the eigensolver converges to the *smallest* eigenvalue λ_1 and to a corresponding eigenvector, since the whole iteration may take place in the orthogonal complement of the invariant subspace to λ_1. But as the latter set forms a null set, and as an effect of rounding errors, such an early breakdown does not occur in practice so that the preconditioned eigensolver converges from scratch. Moreover, note that even inverse iteration is most unstable in the directions of the invariant subspace to λ_1 as all the eigenvectors $x_2, \dots, x_{n-1}$ corresponding to the eigenvalues $\lambda_2, \dots, \lambda_n$ are saddle points of the Rayleigh quotient.

It is an important result that Theorem 2.8 predicts *grid independent convergence* (i.e. there is no dependence on the mesh size h or on the number of the unknowns) for any eigenvalue problem, which stems from the discretization of an elliptic partial differential operator inasmuch γ is bounded away from 1 independently of the mesh size. Then mesh independence of (2.28) follows from the fact that (2.27) is only a function of λ, λ_i, λ_{i+1} and γ. Explicitly, it does not depend on the largest eigenvalue λ_n. As the dependence of the function $\lambda_{i,j}(\lambda, \gamma)$ on its arguments cannot easily be grasped, we refer to Chapter 4 where a simple upper convergence estimate is provided.

As it has already been mentioned, we assume that there is no implicit dependence on λ_n or the mesh size via γ: For the best multigrid or multilevel preconditioners, (2.2) is satisfied for

γ bounded away from 1 independently of the mesh size. Furthermore, in case of an *adaptive* multigrid eigenvalue computation with a good coarse grid approximation, one expects that all the eigenvalue approximations, which are generated in the course of the iteration on all levels of refinement, are located in the interval $[\lambda_1, \lambda_2)$ if the discretization error is small in comparison to $\lambda_2 - \lambda_1$. In this case the bound Θ by (2.29) gives a reliable convergence rate estimate.

Hence, depending on the quality of the preconditioner, *eigenvector/eigenvalue computation can be done with a grid independent rate while the convergence rates are of comparable magnitude with that of multigrid methods for boundary value problems*. Therefore our preconditioned eigensolvers can be viewed as the counterparts of multigrid algorithms for the solution of boundary value problems, see [75, 97] and cf. Table 1.1.

3. Analysis of fastest convergence

In Chapter 2 we have presented *sharp* convergence estimates on the *poorest* convergence of preconditioned inverse iteration. The convergence theory discussed so far appears to have reached some final state as sharpness of these bounds means that they can be attained under the assumptions made within our setup.

Nevertheless, since analytic error estimates on eigenvalue approximations often tend to be pessimistic, the aim of this chapter is to explore the range between *fastest* and *poorest* convergence. To this end we will derive sharp estimates on the best convergence. These estimates from below reveal a wide corridor between fastest and poorest convergence.

Moreover, we will show that the preconditioned eigensolver may converge in a *single step* to an eigenvector belonging to the smallest eigenvalue λ_1. This may happen under relatively weak conditions, i.e. we only assume some *lower* spectral bound on the quality of the preconditioner. We emphasize that such a behavior is fundamentally different from that of inverse iteration! It is well known that inverse iteration (aside from the trivial case that the iteration vector is still an eigenvector) necessarily converges in infinitely many steps. The mentioned properties hold analytically—they have nothing to do with a numerical implementation using finite precision arithmetic. For an illustration of these relations, which however anticipates the analytic results to be derived later within this chapter, the reader may compare the convergence curves for the model problem drawn in Figure 2.2 to those of Figure 3.5 and will observe that extremely fast convergence is possible. Analytic results on this fastest convergence are given in Lemma 3.2 and in Corollary 4.8.

There is a second reason why we are interested in estimates on the fastest convergence: First observe that for the "expensive" choice $B^{-1} = A^{-1}$, which we like to refer as *exact preconditioning*, the convergence estimate of Theorem 2.8 turns into that of inverse iteration (2.32). In other words the convergence rate of inverse iteration appears as the limit rate for exact preconditioning. Moreover, estimate (2.28) for a larger spectral radius of the error propagation matrix, predicts even a poorer convergence estimate compared to that of inverse iteration. Therefore, it is sometimes believed that preconditioned inverse iteration cannot converge faster than inverse iteration. As pointed out above, this is not the case, i.e. the preconditioned eigensolver may *converge significantly faster* than inverse iteration. Here we do not treat the question of how to construct such preconditioners responsible for fastest convergence in practical applications as we restrict the assumption on the preconditioner on the condition (2.2).

The results of the present chapter should also make clear that any high-accuracy solution of the linear system associated with inverse iteration does not pay out. In other words, it is not worthwhile to consider high-accuracy preconditioners for the eigensolvers under consideration. The computational effort necessary for accurate preconditioning should rather be spent to the next step of the eigensolver using a "moderate-quality" preconditioner.

Let us now start with rather technical items. In the following we mainly apply that techniques, which have already been used to prove Theorem 2.8. For given $\gamma \in [0, 1]$ and $\lambda \in (\lambda_1, \lambda_n)$ there are two factors responsible for the speed of convergence:

1. The choice of the preconditioner $B^{-1} \in \mathcal{B}_\gamma$ by (2.3). The most advantageous choice in $\mathcal{B}_\gamma$ consists in a preconditioner which maps the actual iterate x into the (in most cases unique) point of an infimum of the Rayleigh quotient on E_γ. The necessary analysis is given in Section 3.1.

2. The freedom to choose x from the level set

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

   of vectors having the fixed Rayleigh quotient λ. This analysis is presented in Section 3.2.

The choices in $\mathcal{B}_\gamma$ and $L(\lambda)$, optimal in such a way that the Rayleigh quotient of the new PINVIT iterate takes its smallest possible value, defines the vector of fastest convergence

$$x^* \in \arg \min_{x \in L(\lambda)} \min_{B^{-1} \in \mathcal{B}_\gamma} \lambda(x - B^{-1}(Ax - \lambda x)),$$

for which finally, by means of a mini-dimensional analysis, convergence estimates are derived.

Not surprisingly in the light of the discussion above, Theorem 3.15 reveals that PINVIT may converge much more rapidly than expressed by Theorem 2.8. In particular and as mentioned above, one-step convergence to an eigenvector is possible, i.e. eigenvectors are "often" contained in the set E_γ as shown in Lemma 3.2.

3.1 Best preconditioning

For the remaining part of this chapter we employ the c-basis as introduced in Definition 2.5 and which has already been proved to set up the proper geometry for the convergence analysis in [95, 96]. For simplicity, we also assume $c \in \mathbb{R}^n$ to satisfy the following Assumption 3.1.

Assumption 3.1.

1. $\|c\| = 1$*,*

2. c *is not equal to any of the unit vectors* e_i*,* $i = 1, \ldots, n$*,*

3. $c \geq 0$ *componentwise.*

Assumption 3.1 does not mean any restriction of generality. Restricting PINVIT to the unit ball $\|c\| = 1$ is justified since (2.19) is homogeneous with respect to a scaling of the iterate. Excluding the unit vectors, which are the c-basis representations of the eigenvectors of A, avoids stationarity. Finally, for the sake of convenience we restrict the analysis to componentwise nonnegative $c \in \mathbb{R}^n$. Any change of a sign of some component c_k leads to a reflection of $E_\gamma(c)$ by a hyperplane through the origin orthogonal to the kth unit vector. Such a reflection has no effect on the Rayleigh quotient on $E_\gamma(c)$ since $\lambda(\cdot)$ is purely quadratic in the components of its argument.

3.1.1 Extremum points in $E_\gamma(c)$

Lemma 3.2 shows that misconvergence of PINVIT to the largest eigenpair (e_n, λ_n) will never happen and provides a condition under which immediate termination within the first eigenvector e_1 may take place. But let us first introduce the smallest *circular cone* $C_\gamma(c)$ enclosing $E_\gamma(c)$

$$C_\gamma(c) := \{\xi d : \ d \in E_\gamma(c), \ \xi > 0\}. \tag{3.1}$$

Lemma 3.2. *Let* $c \in \mathbb{R}^n$ *obey Assumption 3.1. Then* $e_n \notin C_\gamma(c)$*. Furthermore,* $e_1 \in C_\gamma(c)$ *iff*

$$c_1 \geq \frac{\lambda_1}{\lambda}\Big(\|\lambda\Lambda^{-1}c\|^2 - \gamma^2\|(I - \lambda\Lambda^{-1})c\|^2\Big)^{1/2}. \tag{3.2}$$

Proof. The acute angle χ enclosed between e_n and the axis $\lambda\Lambda^{-1}c$ of $C_\gamma(c)$ is given by

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

Contrastingly, for the opening angle φ of the largest possible cone $C_1(c) \supset C_\gamma(c)$ it holds

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

Since $\|c\| = 1$ and thus $\lambda\lambda_n^{-1}c_n < 1$, we infer that $\varphi < \chi$ which implies $e_n \notin C_\gamma(c)$.

For the acute angle χ between e_1 and $\lambda\Lambda^{-1}c$ it holds

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

and for the opening angle φ of the circular cone $C_\gamma(c)$

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

The condition $\chi \leq \varphi$ reads $\lambda\lambda_1^{-1}c_1 \geq (\|\lambda\Lambda^{-1}c\|^2 - \gamma^2\|(I - \lambda\Lambda^{-1})c\|^2)^{1/2}$. □

Inequality (3.2) is not an unrealistic condition and becomes even weaker for increasing γ. The limit $\gamma \to 1$ defines the largest cone $C_1(c)$ containing e_1, iff

$$c_1 \geq \frac{\lambda_1}{\lambda}, \tag{3.3}$$

which follows from (3.2) together with Lemma 2.4. Therefore, the condition (3.3) will be fulfilled sooner or later whenever c converges to e_1.

Our next aim is to locate points of extrema of the Rayleigh quotient on $E_\gamma(c)$ by analyzing its local behavior. The maximum of $\lambda(E_\gamma(c))$ corresponds to poorest convergence (concerning the choice of the preconditioner in $\mathcal{B}_\gamma$), while the minimum is associated with the best possible convergence. Since the Rayleigh quotient is invariant with respect to nonzero scaling of its argument, we always subsume under uniqueness of points of extrema the uniqueness despite of scaling.

In Lemma 3.3 we recapitulate the gradient and Hessian of the Rayleigh quotient. The maximum and the minimum of $\lambda(\cdot)$ are taken in e_n and e_1, respectively. All the other eigenvectors e_i, $1 < i < n$, are saddle points which can easily be seen by inspecting (3.4) and (3.5).

Lemma 3.3. *For nonzero $c \in \mathbb{R}^n$ the gradient of the Rayleigh quotient (2.18) reads*

$$\nabla\lambda(c) = \frac{2}{(c, \lambda\Lambda^{-1}c)}\,(I - \lambda\Lambda^{-1})c. \tag{3.4}$$

It holds $\nabla\lambda(c) = 0$, iff $c = \theta e_i$ for $1 \leq i \leq n$ and $\theta \neq 0$. The Hessian $H(c)$ of $\lambda(c)$ is given by

$$\begin{aligned} H(c) \;&=\; \frac{2}{(c, \Lambda^{-1}c)^2}\Big[(I - \lambda\Lambda^{-1})(c, \Lambda^{-1}c) \\ &\qquad -2(\Lambda^{-1}c)[(I - \lambda\Lambda^{-1})c]^T - 2[(I - \lambda\Lambda^{-1})c](\Lambda^{-1}c)^T\Big]. \end{aligned} \tag{3.5}$$

Proof. By direct computation from (2.18). (Because of the nonlinearity of (2.18) the Equation (3.4) cannot be gained by applying the linear transformation (2.17) to the x-basis representation of the gradient as given by (1.4). The same does hold concerning the Hessian of the Rayleigh quotient.) □

By using Lemma 3.3 we conclude that the points of absolute extrema of $\lambda(\cdot)$ are located on the boundary $\partial E_\gamma(c)$ of $E_\gamma(c)$. The uninteresting case $e_1 \in C_\gamma(c)$ can be excluded since one-step-convergence is sufficiently described in Lemma 3.2.

Lemma 3.4. *If e_1 is not contained in the interior $\overset{\circ}{C}_\gamma(c)$ of $C_\gamma(c)$ (cf. (3.2)), then*

$$\arg\,\mathrm{ext}\,\lambda(E_\gamma(c)) \subset \partial E_\gamma(c),$$

where ext *denotes the set of absolute extrema.*

Proof. Let w be the point of a minimum or maximum of the Rayleigh quotient on $E_\gamma(c)$ and assume $w \in \overset{\circ}{E}_\gamma(c)$. Then $\nabla\lambda(w) = 0$ and $w = \theta e_i$ by Lemma 3.3 for some scalar θ. Lemma 3.2 guarantees $i \neq n$, while $i \neq 1$ holds by the assumption. In the remaining cases $2 \leq i \leq n-1$ the Hessian (3.5) in θe_i is diagonal,

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

Since $(1 - \lambda_i/\lambda_n) > 0$ and $(1 - \lambda_i/\lambda_1) < 0$ the Hessian has (at least) one positive and one negative eigenvalue, so that w is not a point of an absolute extremum. □

In analogy to Theorem 4.3 in [96], the fact that all points of infima are located on the surface of $E_\gamma(c)$ gives rise to some orthogonal decomposition:

Theorem 3.5. *Let c satisfy Assumption 3.1 and $\gamma \in [0,1)$. If $e_1 \notin C_\gamma(c)$, then it holds for $w \in \arg\inf\lambda(E_\gamma(c))$ that*

$$(w, w - \lambda\Lambda^{-1}c) = 0, \tag{3.6a}$$

$$\|\lambda\Lambda^{-1}c\|^2 = \|w\|^2 + \left\|w - \lambda\Lambda^{-1}c\right\|^2, \tag{3.6b}$$

$$\left\|w - \lambda\Lambda^{-1}c\right\| = \gamma\|(I - \lambda\Lambda^{-1})c\|, \tag{3.6c}$$

$$\|w\| > \|c\|. \tag{3.6d}$$

Proof. If $(w, w - \lambda\Lambda^{-1}c) \neq 0$, then $\kappa w \in \overset{\circ}{E}_\gamma(c)$ with $\kappa = (w, \lambda\Lambda^{-1}c)/(w,w)$ since

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

Additionally, we have $\lambda(\kappa w) = \lambda(w)$ in contradiction to Lemma 3.4. Equation (3.6b) is a direct consequence of (3.6a). Equation (3.6c) only expresses $w \in \partial E_\gamma(c)$. Finally,

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

□

3.1.2 A necessary condition for infimum points

In this section we derive a necessary condition characterizing infimum points of the Rayleigh quotient on $E_\gamma(c)$. This is done by two alternative approaches. First, Lemma 3.6 exploits some orthogonality (3.8) as suggested by Knyazev [67]. Alternatively, in Lemma 3.7 we apply the method of Lagrange multipliers to the Rayleigh quotient and take (3.6a)–(3.6c) as the (geometric) constraints. Obviously, Lemmata 3.6 and 3.7 hold for any constrained local extremum of $\lambda(\cdot)$ on $\partial E_\gamma(c)$.

Lemma 3.6. *Let c obey Assumption 3.1 and $\gamma \in (0,1)$ so that the interior of $E_\gamma(c)$ is nonempty. Then any point of an infimum $w \in \arg\inf \lambda(E_\gamma(c))$ fulfills*

$$(\lambda(w)\Lambda^{-1} + (\eta - 1)I)w = \eta\lambda\Lambda^{-1}c, \tag{3.7}$$

for some real number η.

Proof. If the gradient

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

in $w \in \arg\inf \lambda(E_\gamma(c))$ vanishes, then (3.7) holds trivially for $\eta = 0$. Now assume $\nabla\lambda(w)$ to be nonzero which implies that $w \neq e_i$ for $1 \leq i \leq n$ because of Lemma 3.3 and $w \in \partial E_\gamma(c)$. A necessary condition for an extremum point w of $\lambda(\cdot)$ on $\partial E_\gamma(c)$ reads

$$0 = \lim_{\varepsilon\to 0} \frac{d}{d\varepsilon}\, \lambda(w + \varepsilon z) = (\nabla\lambda(w), z), \tag{3.8}$$

for all z tangent to $E_\gamma(c)$ in w. Hence, $\nabla\lambda(w)$ is orthogonal to the tangent plane of $E_\gamma(c)$ in w. Equivalently, there is a nonzero $\eta \in \mathbb{R}$ so that

$$w - \lambda(w)\Lambda^{-1}w = \eta(w - \lambda\Lambda^{-1}c),$$

which implies (3.7). □

The second approach to derive a necessary condition for an infimum point in w makes use of the Lagrange multiplier method; cf. Lemma 4.4 in [95].

Lemma 3.7. *Under the assumptions of Lemma 3.6 there are constants $\mu, \nu \in \mathbb{R}$ so that*

$$2(\Lambda^{-1} + (\mu + \nu)I)w = \nu\lambda\Lambda^{-1}c, \tag{3.9}$$

Proof. By (3.6b) and (3.6c) it holds

$$\|w\|^2 = \|\lambda\Lambda^{-1}c\|^2 - \gamma^2\|(I - \lambda\Lambda^{-1})c\|^2. \tag{3.10}$$

Therefore $\|w\|$ has a fixed value for given c and γ. Hence, extrema of $(w, \Lambda^{-1}w)$ are taken in the same arguments as those of the Rayleigh quotient $\lambda(w)$. The Lagrange function $L = L(w, \mu, \nu)$ of $(w, \Lambda^{-1}w)$ with respect to the constraints (3.6a) and (3.10) reads

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

where μ and ν are Lagrange multipliers. A vanishing gradient of L in w is the necessary condition for a constrained local extremum and leads to (3.9). □

The conditions (3.7) and (3.9) are essentially the same. First we have to exclude $\eta = 0$ or $\nu = 0$, respectively, since otherwise w could 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 an infimum on $E_\gamma(c)$. We refer to Lemma A.1 in [96] whose arguments can be extended to hold for infimum points. This gives us the justification to assume $\nu \neq 0$ and $\eta \neq 0$ from now on.

The further analysis is based on condition (3.9). Of course, we are interested in finding an explicit representation of w depending on c and the Lagrange multipliers. But to gain this, the diagonal matrix $D := \Lambda^{-1} + (\mu + \nu)I$ in Equation (3.9) has to be inverted. While for the case of suprema, as shown by Theorem 4.8 in [95], D is invertible, this is not always the case if w is an infimum point. In Section 3.1.5 a numerical example is presented. Problems occur if the first component of c equals 0. Lemma 3.8 secures the invertibility of D_{ii} for the remaining components $i > 1$.

Lemma 3.8. *Under the assumption of Lemma 3.6 let $w \in \arg\inf \lambda(E_\gamma(c))$. Then positiveness of c_k implies $w_k > 0$ for $k = 1, \ldots, n$. Moreover,*

$$w_k = \frac{\lambda\nu}{2 + 2\lambda_k(\mu + \nu)}\, c_k > 0. \tag{3.11}$$

Finally, $c_k = 0$ entails $w_k = 0$ but only for $k = 2, \ldots, n$.

Proof. If $c_k \neq 0$, then $\lambda_k^{-1} + \mu + \nu$ and w_k are nonzero by (3.9) which also results in (3.11).

Now assume $w_k < 0$ and define $\bar{w}$ to be equal to w but change the sign of the kth component. Then $\bar{w}$ is closer to the center $\lambda\Lambda^{-1}c$ since

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

Therefore, $\bar{w}$ is located in the interior of $E_\gamma(c)$. But $\lambda(w) = \lambda(\bar{w})$ contradicts Lemma 3.4 so that w_k must be positive.

Next assume $c_k = c_{k'} = 0$ together with $w_k = w_{k'} \neq 0$. Then (3.9) implies

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

so that $\lambda_k = \lambda_{k'}$ or $k = k'$ because all eigenvalues are simple. We conclude that $c_k = 0$ and $w_k \neq 0$ may hold only for a single component.

Now denote by l the smallest index with $c_l > 0$ and let l' be the largest index with $c_{l'} > 0$. We assume $c_k = 0$ and $w_k \neq 0$ for $l < k < l'$. From (3.9) 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 $w_k = 0$.

Next consider $c_m = 0$ and $w_m \neq 0$ for some m with $l' < m \leq n$. Let $\bar{w}$ be equal to w except for the mth component which is set to 0. By construction it holds $\bar{w} \in \mathring{E}_\gamma(c)$. Since $c_m = c_{m+1} = \ldots = c_n = 0$ and thus $\lambda(\lambda\Lambda^{-1}c) < \lambda_m$, we conclude $\lambda(w) = \inf \lambda(E_\gamma(c)) < \lambda_m$. We can rewrite the latter inequality as

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

which is equivalent to $\lambda(\bar{w}) < \lambda(w)$ and which contradicts the assumption of w being an infimum point.

Finally, let $c_m = 0$ and $w_m \neq 0$ for some m with $1 < m < l'$. Define $\bar{w}$ to be equal to w but interchange the components with indexes 1 and m. Since $c_1 = c_m = 0$ one has

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

and thus $\bar{w} \in E_\gamma(c)$. But then due to $\lambda_1 < \lambda_m$ we have $\lambda(\bar{w}) < \lambda(w)$, which contradicts $w \in \arg\inf \lambda(E_\gamma(c))$. □

3.1.3 Parametrization of infimum points

Let us now assume $c_1 \neq 0$. This assumption, on the one hand, allows us to show that for any γ the infimum point on $E_\gamma(c)$ is unique. On the other hand, we are able to parametrize the continuous curve of infima for $\gamma \in [0,1)$ in some real parameter α.

Theorem 3.9. *On the assumptions of Lemma 3.6 and assuming that $c_1 > 0$ any infimum point $w \in \arg\inf \lambda(E_\gamma(c))$ can be written as*

$$w = \beta(\alpha I + \Lambda)^{-1}c \tag{3.12}$$

for unique real numbers $\alpha \in (-\lambda_1, 0]$ and

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

Moreover, the function

$$\rho : (-\lambda_1, 0] \to (\lambda_1, \lambda(\Lambda^{-1}c)] : \alpha \mapsto \lambda(w) = \lambda((\alpha I + \Lambda)^{-1}c) \tag{3.13}$$

is strictly monotone increasing in α. Finally, the gradient vector in w and $w - \lambda\Lambda^{-1}c$, the normal vector on $E_\gamma(c)$ in w, are collinear. Thus

$$\nabla\lambda(w) \in \operatorname{span}\{w, \lambda\Lambda^{-1}c\}. \tag{3.14}$$

Proof. Summarizing the results of Lemmata 3.7 and 3.8, any $w \in \arg\inf \lambda(E_\gamma(c))$ can be written in the form (3.12) for $\alpha, \beta \in \mathbb{R}$. The coefficients α and β are functions of $\gamma \in [0, 1)$.

First we show that $\beta > 0$ and $\alpha > -\lambda_1$: For $w = \beta(\alpha I + \Lambda)^{-1}c$ we have $\beta/(\alpha+\lambda_i) > 0$ for any nonzero component c_i by Lemma 3.8. 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, $\lambda(w) > \lambda(c)$, which contradicts PINVIT convergence, see Theorem 2.8. The explicit form of $\beta > 0$ can be gained from (3.6a).

In order to show that ρ is strictly monotone increasing, 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 $i = 1, \dots, n$

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

Therein the positive coefficients $(\alpha_2 + \lambda_1)/(\alpha_1 + \lambda_1), \dots, (\alpha_2 + \lambda_n)/(\alpha_1 + \lambda_n)$ are strictly monotone decreasing. Applying Lemma A.1 in [95] shows that ρ is strictly monotone increasing. Furthermore, it holds

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

Uniqueness of α follows, since in the remaining case $\alpha > 0$ we would have $\lambda((\alpha I + \Lambda)^{-1}c) > \lambda(\Lambda^{-1}c)$, which is impossible for $(\alpha I + \Lambda)^{-1}c$ as an infimum point.

Collinearity of $\nabla\lambda(w)$ and $w - \lambda\Lambda^{-1}c$ is only a reformulation of (3.8) which immediately results in (3.14). □

3.1.4 Dependence of the shift parameter α on γ

Looking back one can say that the important and somewhat surprising result of Theorem 3.9 is a representation of points of *infima* of the Rayleigh quotient on $E_\gamma(c)$ within some real parameters α and β. Its counterpart concerning the representation of points of *suprema* has been given in [96] and leads to a formula like (3.12), too. Since the Rayleigh quotient is invariant with respect to the choice of β, we conclude that *extrema on $E_\gamma(c)$ can be represented by using only the single real control parameter α.*

Consequently, the next challenging problem is to derive an explicit formula describing the dependence of α on γ, which would then allow a convenient and useful representation of the extremal Rayleigh quotients on $E_\gamma(c)$ only depending on γ and the vector c. Without doubt, such an approach could simplify parts of the PINVIT convergence theory considerably! First of all, it would allow us to analyze the dependence of the extrema of the Rayleigh quotient $\lambda(w[c])$ with respect to all vectors c having a fixed Rayleigh quotient, which is a central step in order to find an estimate describing the poorest PINVIT convergence.

Unfortunately, the problem of finding α in (3.12) as an explicit function of γ is not easy to tackle. We have not succeeded in deriving $\alpha(\gamma)$ in the $\mathbb{R}^n$, $n > 2$, as the determination of $\alpha(\gamma)$ in (3.16) requires the solution of a polynomial with degree $2n-2$ in α. But even the result in the $\mathbb{R}^2$ is of some importance since it provides the basis for deriving simplified PINVIT convergence estimates in Section 4.2.2. Deriving a general formula for $\alpha(\gamma)$ remains to be an open problem.

Let us now determine α and β in $\mathbb{R}^2$: Exploiting the geometry described in Theorem 3.5, we can derive the scaling parameter β in (3.12) by minimizing the distance of $w[\beta]$ from $\lambda\Lambda^{-1}c$ with respect to a variation of β. Then we obtain

$$w = \frac{(\tilde{w}, \lambda\Lambda^{-1}c)}{(\tilde{w}, \tilde{w})}\tilde{w} \tag{3.15}$$

for $\tilde{w} = (\alpha I + \Lambda)^{-1}c$. By using (3.6b) and (3.6c) we obtain

$$(\tilde{w}, \lambda\Lambda^{-1}c)^2 = \|\tilde{w}\|^2 \left(\left\|\lambda\Lambda^{-1}c\right\|^2 - \gamma^2 \left\|(I - \lambda\Lambda^{-1})c\right\|^2 \right) \tag{3.16}$$

from which the function $\gamma(\alpha)$ can easily be determined. But we are interested in the inverse function $\alpha(\gamma)$ which, unfortunately, is a polynomial of the degree $2n-2$ in α.

Let us solve this polynomial for $n = 2$. Then $c \in \mathbb{R}^2$ and $\|c\| = 1$ together with $\lambda_1 < \lambda(c) = \lambda < \lambda_2$ leads to

$$c_1^2 = \frac{\lambda_1(\lambda_2 - \lambda)}{\lambda(\lambda_2 - \lambda_1)} \quad \text{and} \quad c_2^2 = \frac{\lambda_2(\lambda - \lambda_1)}{\lambda(\lambda_2 - \lambda_1)}. \tag{3.17}$$

Hence,

$$\|\lambda\Lambda^{-1}c\|^2 - \gamma^2\|(I - \lambda\Lambda^{-1})c\|^2 = \frac{(\lambda_1 + \lambda_2 - \lambda)\lambda - \gamma^2(\lambda - \lambda_1)(\lambda_2 - \lambda)}{\lambda_1\lambda_2}.$$

Finally, one derives from (3.16)

$$\alpha^\pm = \frac{\gamma\sqrt{\lambda_1\lambda_2}}{\lambda(1-\gamma^2)} \left(\gamma\sqrt{\lambda_1\lambda_2} \pm \sqrt{(1-\gamma^2)(\lambda_2 - \lambda)(\lambda - \lambda_1) + \lambda_1\lambda_2} \right), \tag{3.18}$$

where the negative sign ($\alpha^- < 0$) defines the infimum point of $\lambda(\cdot)$ on $E_\gamma(c)$. Beyond that, we remark that the positive sign ($\alpha^+ > 0$) corresponds to the supremum of the Rayleigh quotient on $E_\gamma(c)$. Both quantities are next used in Section 4.2.2 for deriving concise PINVIT estimates.

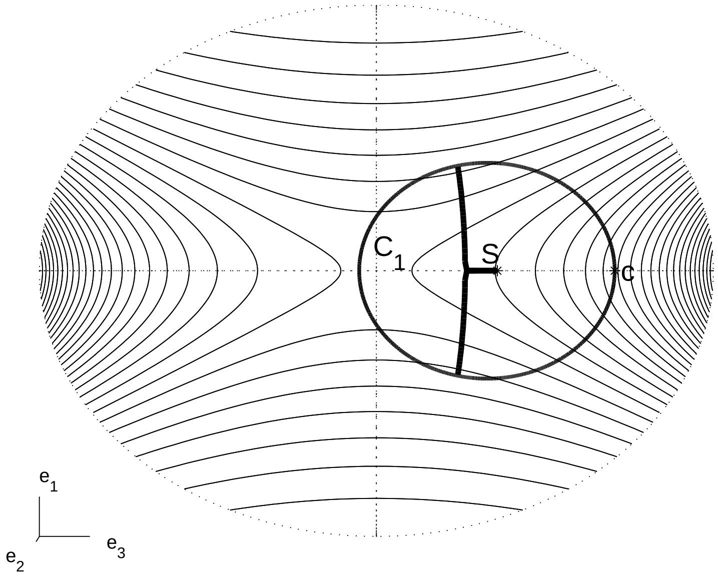


Figure 3.1: *Curve S of infimum points on $E_\gamma(c)$, $\gamma \in [0,1]$, projected on the unit sphere. The bold circle C_1 marks the intersection of $C_1(c)$ with the unit sphere.*

3.1.5 Bifurcation of the infima curve

In order to parametrize the curve of infimum points, we have assumed $c_1 \neq 0$ in Section 3.1.3. This assumption is fulfilled whenever $\lambda(c) < \lambda_2$, as assumed in the classical convergence analysis of preconditioned gradient methods [31, 36].

But if $c_1 = 0$, then Equation (3.13) is not capable of presenting all infimum points, since then $\min \rho(\alpha) = \lambda_2$, but at the same time it may hold $\min \lambda(E_\gamma(c)) < \lambda_2$. A continuity argument can help to understand what happens for $c_1 \to 0$. One finds that as long as

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

where $+$ denotes the pseudoinverse, the form of the infimum points is determined by Theorem 3.9. Beyond the bound $\lambda((\Lambda - \lambda_1 I)^+ c)$ the infimum points have the form (aside from scaling)

$$\pm \vartheta e_1 + (\Lambda - \lambda_1 I)^+ c \tag{3.19}$$

for suitable $\vartheta \geq 0$.

A numerical example for the smallest nontrivial dimension is given in Figure 3.1 for 3 eigenvalues of the test problem (2.30), namely $\Lambda = \mathrm{diag}(2, 5, 13)$. 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 $i = 1, \ldots, 29$.

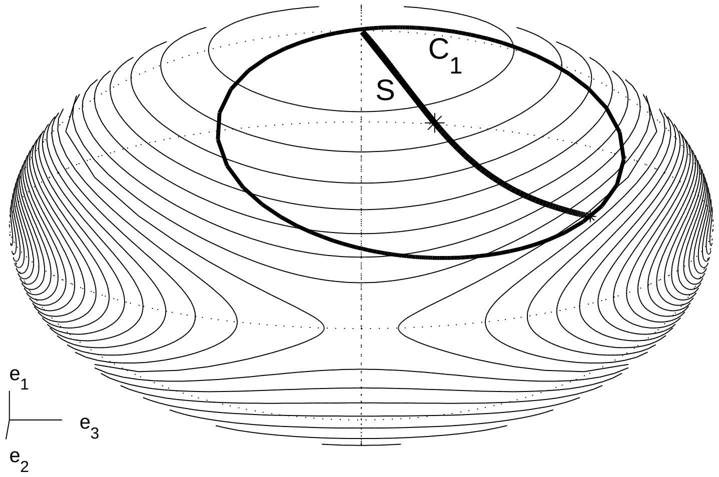


Figure 3.2: *Curve S of absolute extrema on $E_\gamma(c)$,$\gamma \in [0,1]$, projected on the unit sphere.*

For $c = (0, 1/\sqrt{2}, 1/\sqrt{2})^T$ the intersection of $C_1(c)$ with the unit sphere is the bold circle C_1. The curve S of infimum 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 remaining part of the curves is of the form (3.19).

Let us also illustrate in Figure 3.2 the curve S of all absolute extrema for $\gamma \in [0,1]$. This is done for the same Λ but $c = (3,5,5)^T/\sqrt{59}$. Now, α is contained in the interval $(-\lambda_1, \infty)$ and the continuous curve

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

starts at the north pole ($\alpha \to -\lambda_1$), runs through the axis of the cone $C_\gamma(c)$ for $\alpha = 0$ and ends in the initial vector c for $\alpha \to \infty$. Therein, all $\alpha < 0$ correspond to infimum points while $\alpha > 0$ gives the representation of supremum points. For this example the condition (3.3) is fulfilled since $0.391 \approx c_1 > \lambda_1/\lambda \approx 0.387$. Hence, the eigenvector e_1 is contained in $C_1(c)$, but close to its boundary.

3.2 Extremal quantities on $L(\lambda)$

The aim of this section is to describe how the mapping

$$c \to w \in \arg\inf \lambda(E_\gamma(c))$$

depends on the choice of c from the level set

$$L(\lambda) = \{c \in \mathbb{R}^n : \lambda(c) = \lambda, \ \ \|c\| = 1\}. \tag{3.20}$$

Throughout this section we make use of the non-restrictive Assumption 3.1 and assume each eigenvalue to have the multiplicity 1, cf. Section 2.3. Here, we generalize some results gained in Section 2 of [96] from suprema to the case of global extrema.

3.2.1 Extrema of $\|\nabla\lambda(c)\|$

Theorem 3.10 shows that extrema of $\|\nabla\lambda(c)\|$ are taken in 2D (1D) invariant subspaces.

Theorem 3.10. *Let $\lambda = \lambda(c) \in (\lambda_1, \lambda_n)$. Then for the Euclidean norm of the gradient (3.4) on $L(\lambda)$ by (3.20) it holds:*

1. *Minima of $\|\nabla\lambda(c)\|$ are taken either in $c = e_i$ (this uninteresting case is excluded by Assumption 3.1 since then $\lambda = \lambda_i$ and $\nabla\lambda(e_i) = 0$) or for $\lambda_i < \lambda < \lambda_{i+1}$ in*

$$c_{i,i+1} := (0, \dots, 0, c_i, c_{i+1}, 0, \dots, 0)^T \in L(\lambda), \tag{3.21}$$

   having exactly the two non-zero components c_i and c_{i+1}.

2. *Maxima of $\|\nabla\lambda(c)\|$ are all taken in $c_{1,n} = (c_1, 0, \dots, 0, c_n)$.*

Proof. The method of Lagrange multipliers provides a necessary condition for relative extrema of $\|\nabla\lambda(c)\|$ with respect to the constraints $\|c\| = 1$ and $\lambda(c) = \lambda$. Inserting the constraints and squaring the objective functional $\|\nabla\lambda(c)\|$ leads to the Lagrange function

$$L(c) = \left\|(I - \lambda\Lambda^{-1})c\right\|^2 + \mu(\|c\|^2 - 1) + \nu((c, \Lambda^{-1}c) - \lambda^{-1}). \tag{3.22}$$

where μ and ν denote Lagrange multipliers. Hence, $\nabla L = 0$ reads

$$(I - \lambda\Lambda^{-1})^2 c + \mu c + \nu\Lambda^{-1}c = 0. \tag{3.23}$$

If c is not collinear to any of the unit vectors, then there are at least two nonzero components c_k and c_l, $k \neq l$, and (3.23) results in a system of linear equations for μ and ν

$$\begin{pmatrix} 1 & \lambda_k^{-1} \\ 1 & \lambda_l^{-1} \end{pmatrix} \begin{pmatrix} \mu \\ \nu \end{pmatrix} = \begin{pmatrix} -(1 - \lambda\lambda_k^{-1})^2 \\ -(1 - \lambda\lambda_l^{-1})^2 \end{pmatrix},$$

whose determinant does not vanish. Its unique solution reads

$$\mu = \frac{\lambda^2}{\lambda_k\lambda_l} - 1 \quad \text{and} \quad \nu = \frac{\lambda(2\lambda_k\lambda_l - \lambda(\lambda_k + \lambda_l))}{\lambda_k\lambda_l}.$$

Inserting μ and ν in (3.23) we obtain for $j = k, l$ that $\lambda^2(\lambda_j - \lambda_k)(\lambda_j - \lambda_l)(\lambda_j^2 \lambda_k \lambda_l)^{-1} c_j = 0$. Hence, $c_j = 0$ for $j \neq k, l$ and it holds $c = c_{k,l}$ as well as $\lambda \in (\lambda_k, \lambda_l)$. The constraints $\|c\| = 1$ and $\lambda(c) = \lambda$ give

$$c_k^2 = \frac{\lambda_k(\lambda_l - \lambda)}{\lambda(\lambda_l - \lambda_k)} \quad \text{and} \quad c_l^2 = \frac{\lambda_l(\lambda - \lambda_k)}{\lambda(\lambda_l - \lambda_k)},$$

together with

$$\|\nabla\lambda(c)\|^2 = \frac{4}{(c, \Lambda^{-1}c)^2} \left\|(I - \lambda\Lambda^{-1})c\right\|^2 = \frac{4\lambda^2(\lambda - \lambda_k)(\lambda_l - \lambda)}{\lambda_k \lambda_l}. \tag{3.24}$$

Since $\lambda_k < \lambda < \lambda_l$ one finally obtains

$$\frac{d}{d\lambda_k} \|\nabla\lambda(c)\|^2 = -\frac{4\lambda^3(\lambda_l - \lambda)}{\lambda_l \lambda_k^2} < 0 \quad \text{and} \quad \frac{d}{d\lambda_l} \|\nabla\lambda(c)\|^2 = \frac{4\lambda^3(\lambda - \lambda_k)}{\lambda_k \lambda_l^2} > 0.$$

Thus $\|\nabla\lambda(c)\|$ takes its minimum in $c_{i,i+1}$ and its maximum in $c_{1,n}$. □

3.2.2 Extremal properties of the cone $C_\gamma(c)$

The *opening angle* $\varphi_\gamma(c)$ of the circular cone $C_\gamma(c)$,

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

enclosing $E_\gamma(c)$ is defined by

$$\varphi_\gamma(c) := \sup_{z \in C_\gamma(c)} \arccos\left(\frac{\lambda\Lambda^{-1}c}{\|\lambda\Lambda^{-1}c\|}, \frac{z}{\|z\|}\right). \tag{3.25}$$

The complementary angle to $\varphi_\gamma(c)$ in $C_1(c)$ is the *shrinking angle*

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

The shrinking angle turns out to be relevant in the following. We note that the action of PINVIT can be understood as a shrinking of the initial cone $C_1(c)$: while c is located on the surface of $C_1(c)$, all global extrema of the Rayleigh quotient (as long as e_1 is not contained in $C_\gamma(c)$) are taken on the surface of $C_\gamma(c)$. Lemma 3.11 reveals a close relation between $\|\nabla\lambda(c)\|$ and $\varphi_\gamma(c)$, $\psi_\gamma(c)$.

Lemma 3.11. *Let $\lambda \in (\lambda_1, \lambda_n)$ and $\gamma \in [0, 1]$. The trivial cases $\varphi_\gamma(c) = 0$ ($\psi_\gamma(c) = 0$) can only be taken iff $\gamma = 0$ ($\gamma = 1$) or $c = e_i$ for $i = 2, \ldots, n-1$.*

For non-vanishing angles and $\lambda \in (\lambda_i, \lambda_{i+1})$ the opening angle $\varphi_\gamma(c)$ and the shrinking angle $\psi_\gamma(c)$ on the level set $L(\lambda)$ take their minimum in $c_{i,i+1}$ and their maximum in $c_{1,n}$.

Proof. Using Theorem 3.5 we can rewrite $\varphi_\gamma(c)$ as

$$\varphi_\gamma(c) = \arcsin \frac{\gamma \, \|(I - \lambda\Lambda^{-1})c\|}{\|\lambda\Lambda^{-1}c\|} \quad \text{and} \quad \varphi_1(c) = \arctan \frac{\|(I - \lambda\Lambda^{-1})c\|}{\|c\|}. \tag{3.26}$$

In order to show that the proposition holds for $\gamma = 1$, note that $\arctan(\cdot)$ is strictly monotone increasing. Hence, it suffices to check the extremal properties for $\|(I - \lambda\Lambda^{-1})c\| \, / \, \|c\|$. Since λ and $\|c\|$ are fixed on $L(\lambda)$, the opening angle $\varphi_1(c)$ takes its extrema in the same arguments as $\|\nabla\lambda(c)\|$. Therefore, Theorem 3.10 proves the proposition.

Now let $\gamma \in (0,1)$, then we have from (3.26)

$$\sin(\varphi_\gamma(c)) = \gamma \sin(\varphi_1(c)). \tag{3.27}$$

Since $\sin(\cdot)$ is a strictly monotone increasing function on $[0, \frac{\pi}{2}]$ and by using the definitions $\varphi_1(c_{1,n}) := \min\{\varphi_1(c) : \; c \in L(\lambda)\}$ as well as $\varphi_1(c_{i,i+1}) := \max\{\varphi_1(c) : \; c \in L(\lambda)\}$, one obtains for $\gamma \in (0,1)$

$$\begin{aligned} \gamma \sin(\varphi_1(c_{1,n})) &= \min\{\gamma \sin(\varphi_1(c)) : \; c \in L(\lambda)\}, \\ \gamma \sin(\varphi_1(c_{i,i+1})) &= \max\{\gamma \sin(\varphi_1(c)) : \; c \in L(\lambda)\}. \end{aligned}$$

Applying (3.27) as well as the monotonicity of $\sin(\cdot)$ leads to

$$\varphi_\gamma(c_{1,n}) = \min\{\varphi_\gamma(c) : \; c \in L(\lambda)\}, \qquad \varphi_\gamma(c_{i,i+1}) = \max\{\varphi_\gamma(c) : \; c \in L(\lambda)\},$$

which establishes the required result.

To prove the proposition for $\psi_\gamma(c)$, let $a := \|(I - \lambda\Lambda^{-1})c\| \, / \, \|\lambda\Lambda^{-1}c\|$, then

$$\Psi_\gamma(a) := \psi_\gamma(c) = \arcsin(a) - \arcsin(\gamma a).$$

Note that $\varphi_1(c)$ takes its extrema in the same arguments as $a = \sin\varphi_1$. Differentiation of $\Psi_\gamma(a)$ shows that for $\gamma \in (0,1)$

$$\frac{d}{da}\Psi_\gamma(a) = \frac{\sqrt{1-\gamma^2a^2} - \sqrt{1-a^2}}{\sqrt{(1-a^2)(1-\gamma^2a^2)}} > 0.$$

Hence, $\Psi_\gamma(a)$ is strictly monotone increasing in a which completes the proof. □

3.2.3 Angle dependence of the Rayleigh quotient on $C_\gamma(c)$

To analyze the dependence of the Rayleigh quotient on the opening angle φ_γ on $C_\gamma(c)$ we define the plane

$$P_{c,w} := \operatorname{span}\{\lambda\Lambda^{-1}c, w\}, \tag{3.28}$$

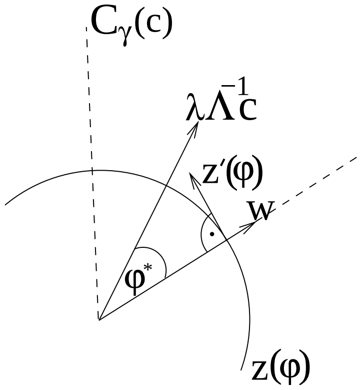


Figure 3.3: *The 2D cross-section $P_{c,w}$.*

where w given by (3.12) denotes a point of an infimum of the Rayleigh quotient on $C_\gamma(c)$. Now parametrize the unit circle in $P_{c,w}$ by $z(\varphi)$ so that $\varphi = \measuredangle(z(\varphi), \lambda\Lambda^{-1}c)$ and $z(\varphi^*) = w/\left\|w\right\|$ for $\varphi^* < \pi$.

To express the angle dependence of the Rayleigh quotient in $P_{c,w}$ we define

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

Lemma 3.12. *On the assumptions of Theorem 3.9 let φ^* so that $z(\varphi^*) = w/\left\|w\right\|$. Then it holds*

$$\left|\frac{d\lambda_{c,w}}{d\varphi}(\varphi^*)\right| = \|\nabla\lambda(\frac{w}{\|w\|})\|. \tag{3.29}$$

Proof. The chain rule yields

$$\frac{d}{d\varphi}\lambda(z(\varphi)) = (\nabla\lambda(z(\varphi)), z'(\varphi)). \tag{3.30}$$

Since $\|z(\varphi)\| = 1$, the derivative $z'(\varphi)$ with $\|z'(\varphi)\| = 1$ is tangent to the unit circle in $P_{c,w}$, i.e. $(z(\varphi), z'(\varphi)) = 0$. By (3.14) the gradient $\nabla\lambda(w/\left\|w\right\|)$ is contained in $P_{c,w}$ and is collinear to the tangent vector $z'(\varphi^*)$ in $w/\left\|w\right\|$, cf. Figure 3.3. We conclude

$$z'(\varphi^*) = \pm\frac{\nabla\lambda(v)}{\|\nabla\lambda(v)\|}. \tag{3.31}$$

Inserting (3.31) in (3.30) for $\varphi = \varphi^*$ completes the proof. □

Now define by $\underline{\lambda}(c, \varphi)$ 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 3.13 discloses the identity of the derivatives $(d\bar{\lambda}(c,\varphi)/d\varphi)$ and $(d\lambda_{c,w}(\varphi)/d\varphi)$ within infimum points.

Lemma 3.13. *On the assumptions of Theorem 3.9 let w be an infimum 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*

$$|\frac{d\underline{\lambda}}{d\varphi}(c,\varphi^*)| = |\frac{d\lambda_{c,w}}{d\varphi}(\varphi^*)| = \|\nabla\lambda(\frac{w}{\|w\|})\|. \tag{3.32}$$

Proof. Both $\lambda_{c,w}(\varphi)$ and $\bar{\lambda}(c,\varphi)$ 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 infimum point $w/\,\|w\|$. Since $\lambda_{c,w}(\varphi) - \underline{\lambda}(c,\varphi)$ is a positive differentiable function taking its minimum in φ^*, we conclude

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

The proposition follows with (3.29). □

3.3 Mini-dimensional convergence analysis

In Section 3.2 we have learnt that several quantities which define the geometry of PINVIT take their extremal values in 2D invariant subspaces. Hence, not surprisingly, PINVIT takes its extremal convergence exactly in these 2D invariant subspaces. To pave the way for the PINVIT convergence Theorem 3.15, we now carry out a mini-dimensional analysis in $\operatorname{span}\{e_i, e_j\}$. Therefore, we pursue an alternative approach compared to the construction used in Theorem 5.1 in [95].

Theorem 3.14. *Let $c \in \mathbb{R}^2$, $\|c\| = 1$ and $\Lambda = \operatorname{diag}(\lambda_j, \lambda_i)$, $\lambda_i < \lambda_j$ (reversed order of eigenvalues). The Rayleigh quotient in the point of a supremum $w_1 \in \arg\sup \lambda(E_\gamma(c))$ reads*

$$\lambda(w_1) = \lambda_{i,j}(\lambda, \gamma, m_1), \tag{3.33}$$

and whenever $e_i \notin C_\gamma(c)$ and $w_2 \in \arg\inf \lambda(E_\gamma(c))$ it holds

$$\lambda(w_2) = \lambda_{i,j}(\lambda, \gamma, m_2), \tag{3.34}$$

where

$$\lambda_{i,j}(\lambda, \gamma, m) = \lambda_i\lambda_j \left(\lambda_j - \frac{\lambda_j - \lambda_i}{1+m^2}\right)^{-1}. \tag{3.35}$$

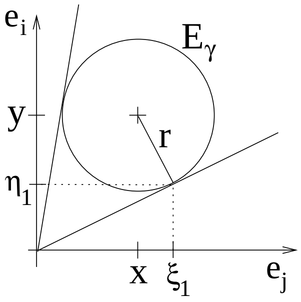


Figure 3.4: *Geometric setup in* $\operatorname{span}\{e_i, e_j\}$.

Therein m_1 *(*m_2*) is the slope of a ray through 0 and tangent to* $E_\gamma(c)$ *which maximizes (minimizes) the Rayleigh quotient with*

$$m_1 = \frac{yl - rx}{xl + ry}, \quad \textit{and} \quad m_2 = \frac{yl + rx}{xl - ry}, \tag{3.36}$$

where $(x, y)^T = \lambda\Lambda^{-1}c$*,* $r = \gamma\|(I - \lambda\Lambda^{-1})c\|$ *and* $l = \sqrt{x^2 + y^2 - r^2}$*.*

One explicitly obtains $\lambda^+_{i,j} = \lambda_{i,j}(\lambda, \gamma, m_1)$ *and* $\lambda^-_{i,j} = \lambda_{i,j}(\lambda, \gamma, m_2)$ *in the form*

$$\begin{aligned} \lambda^{\pm}_{i,j}(\lambda, \gamma) \quad := \quad & \lambda\lambda_i\lambda_j(\lambda_i + \lambda_j - \lambda)^2 / \\ & \Big(\gamma^2(\lambda_j - \lambda)(\lambda - \lambda_i)(\lambda\lambda_j + \lambda\lambda_i - \lambda_i^2 - \lambda_j^2) \\ & \mp 2\gamma\sqrt{\lambda_i\lambda_j}(\lambda - \lambda_i)(\lambda_j - \lambda) \\ & \qquad \times\sqrt{\lambda_i\lambda_j + (1 - \gamma^2)(\lambda - \lambda_i)(\lambda_j - \lambda)} \\ & -\lambda(\lambda_i + \lambda_j - \lambda)(\lambda\lambda_j + \lambda\lambda_i - \lambda_i^2 - \lambda_i\lambda_j - \lambda_j^2)\Big). \end{aligned} \tag{3.37}$$

Proof. Since all global extrema of $\lambda(\cdot)$ on $E_\gamma(c)$ are located on the surface of the enclosing cone $C_\gamma(c)$, we compute the two points of intersection of $\partial E_\gamma(c)$ and $\partial C_\gamma(c)$. Therefore, consider the following parametrization of the circle $\partial E_\gamma(c)$ by

$$E(\varphi) = \lambda\Lambda^{-1}c + r\begin{pmatrix} \sin\varphi \\ \cos\varphi \end{pmatrix}, \qquad \varphi \in [0, 2\pi).$$

We are looking for all m so that the ray $(\xi, m\xi)$, $\xi \in \mathbb{R}$, has a unique point of intersection with $E_\gamma(c)$, see Figure 3.4. Elimination of ξ results in

$$m(x + r\sin\varphi) = y + r\cos\varphi,$$

so that we are looking for the minimum m_1 and maximum m_2 of

$$g(\varphi) := \frac{y + r\cos(\varphi)}{x + r\sin(\varphi)}.$$

The necessary condition $(dg(\varphi)/d\varphi) = 0$ reads

$$x\sin(\varphi) + y\cos(\varphi) + r = 0.$$

From this we determine m_1 and m_2 as given by (3.36) and obtain as the points of intersection

$$(\xi_1, \eta_1) = (\frac{xl^2 + ryl}{x^2 + y^2}, \sqrt{l^2 - \xi_1^2}), \tag{3.38}$$

$$(\xi_2, \eta_2) = (\frac{xl^2 - ryl}{x^2 + y^2}, \sqrt{l^2 - \xi_2^2}). \tag{3.39}$$

The Rayleigh quotient (2.18) of $(\xi, \eta)^T$ reads

$$\lambda_{i,j}^{\pm}(\lambda, \gamma) = \lambda((\xi_{1,2}, \eta_{1,2})^T) = \lambda_i\lambda_j \left(\lambda_j - \frac{\lambda_j - \lambda_i}{1 + m_{1,2}^2} \right)^{-1}.$$

In order to derive (3.37), we determine the components of the positive vector $c \in \mathbb{R}^2$ with $||c|| = 1$ and $\lambda(c) = \lambda$

$$c_i = \left(\frac{\lambda_i(\lambda_j - \lambda)}{\lambda(\lambda_j - \lambda_i)} \right)^{1/2}, \qquad c_j = \left(\frac{\lambda_j(\lambda - \lambda_i)}{\lambda(\lambda_j - \lambda_i)} \right)^{1/2}. \tag{3.40}$$

Hence $(x, y)^T = \lambda\Lambda^{-1}c$ reads

$$x = \sqrt{\frac{\lambda(\lambda - \lambda_i)}{\lambda_j(\lambda_j - \lambda_i)}}, \qquad y = \sqrt{\frac{\lambda(\lambda_j - \lambda)}{\lambda_i(\lambda_j - \lambda_i)}}, \tag{3.41}$$

while r and l are given by

$$r = \gamma\sqrt{\frac{(\lambda - \lambda_i)(\lambda_j - \lambda)}{\lambda_i\lambda_j}}, \quad l = \sqrt{\frac{\gamma^2(\lambda_i - \lambda)(\lambda_j - \lambda) + \lambda(\lambda_i + \lambda_j - \lambda)}{\lambda_i\lambda_j}}. \tag{3.42}$$

Inserting (3.41) and (3.42) in (3.36) and (3.35) results, after wearisome calculations, in (3.37). □

3.4 Convergence estimates for the Rayleigh quotient

While in [95, 96] only the *poorest* convergence of PINVIT(1) is analyzed, we now collect the results gained in the previous Sections 3.1–3.3 to formulate estimates concerning its *best* (or fastest possible) convergence. We differentiate between the best and poorest choice of both the preconditioner and also that of the best and poorest choice of the iteration vector from the level set $L(\lambda)$. Clearly the estimates are sharp since we have (implicitly) constructed the preconditioners of best/poorest convergence as well as the vector c in which best/poorest convergence is attained. The reader should be aware of the fact that all estimates hold independently of the choice of the basis (c- or x-basis).

Theorem 3.15 (PINVIT convergence estimates on fastest convergence). *Consider the level set $L(\lambda)$ of coefficient vectors $c \in \mathbb{R}^n$ with respect to the c-basis having the Rayleigh quotient $\lambda = \lambda(c) \in (\lambda_i, \lambda_{i+1})$ for some index i, $1 \leq i < n$. Furthermore, we consider all admissible preconditioners satisfying the constraint (2.7) for a given $\gamma \in [0, 1)$.*

Then for the fastest convergence of preconditioned inverse iteration (depending on the choice of the preconditioners from the set of all admissible preconditioners as well as for all $c \in L(\lambda)$, i.e. the level set of all vectors having the Rayleigh quotient λ) it holds that:

1. *If $e_1 \in C_\gamma(c)$, i.e. the infimum of the Rayleigh quotient on the set of possible iterates $E_\gamma(c)$ equals λ_1, then PINVIT may terminate in a single step within an eigenvector corresponding to the smallest eigenvalue λ_1. (A necessary and sufficient condition describing this case of immediate termination is given by Lemma 3.2.)*

   Otherwise, the largest possible decrease of the Rayleigh quotient within one step of PINVIT (for the most favorable choice of the preconditioner and the best choice $c \in L(\lambda)$) is attained in $c = c_{1,n}$ with

   $$c_{1,n} = (c_1, 0, \ldots, 0, c_n).$$

   The associated smallest possible Rayleigh quotient λ' of the new iterate is given by

   $$\lambda' = \lambda^-_{1,n}(\lambda, \gamma),$$

   where $\lambda^-_{1,n}(\lambda, \gamma)$ is defined by Equation (3.37).

2. *Under the assumption of 1., one obtains for the choice of the poorest preconditioner (maximizing the Rayleigh quotient on $E_\gamma(c)$) if applied to the vector of fastest convergence $c_{1,n} \in L(\lambda)$ as the Rayleigh quotient of the new PINVIT iterate*

   $$\lambda' = \lambda^+_{1,n}(\lambda, \gamma).$$

3. *Once more assume $e_1 \notin C_\gamma(c)$. Then best preconditioning within the vector of poorest convergence $c_{i,i+1} \in L(\lambda)$ (compare Theorem 2.8) results in the Rayleigh quotient*

$$\lambda' = \lambda^-_{i,i+1}(\lambda, \gamma)$$

as an upper estimate for the fastest decrease of the Rayleigh quotient.

Proof. To show (1), our idea is to follow the curve of extremum points (infima and suprema) as given by

$$(\alpha I + \Lambda)^{-1} c / \left\|(\alpha I + \Lambda)^{-1} c\right\|,$$

for $\alpha \in (\alpha_{\min}, \infty)$ with a proper choice of $\alpha_{\min}$, and to compare the decrease of the Rayleigh quotient along this curve with that on the analogous curve of extremum points defined by $c_{1,n} \in L(\lambda)$. We call these curves $S(c)$ and $S(c_{1,n})$. We start on these curves at c and $c_{1,n}$ (having equal Rayleigh quotients), run along $S(c)$, ($S(c_{1,n})$), and finally end in the infimum points on $C_\gamma(c)$, ($C_\gamma(c_{1,n})$).

First note that by Lemma 3.11 the opening angle φ_γ takes its maximum on $L(\lambda)$ in $c_{1,n}$ so that for any $c \in L(\lambda)$

$$\varphi_\gamma(c_{1,n}) \geq \varphi_\gamma(c), \qquad \gamma \in [0, 1].$$

Let v ($v_{1,n}$) be two normed extremum points (either infima or suprema) on the curves S ($S(c_{1,n})$) in such a way that $\lambda(v) = \lambda(v_{1,n})$. The angles enclosed with the axes of the associate cones are denoted by

$$\varphi^* = \angle(v, \lambda\Lambda^{-1}c) \quad \text{and} \quad \varphi^*_{1,n} = \angle(v_{1,n}, \lambda\Lambda^{-1}c_{1,n}).$$

We parametrize the curves S in φ starting at c ($c_{1,n}$) for $\varphi = 0$ and ending in

$$\varphi^*_{1,n} + \varphi_\gamma(c_{1,n}) \geq \varphi^* + \varphi_\gamma(c),$$

if v ($v_{1,n}$) are infimum points. As long as v ($v_{1,n}$) are supremum points, we only consider the complementary shrinking angles, see the proof of Theorem 1.1 in [96].

Let $\lambda(c, \varphi)$ be the Rayleigh quotient on the curve $S(c)$ as parametrized in φ. Then for the derivatives of $\lambda(c, \varphi)$ and $\lambda(c_{1,n}, \tilde{\varphi}_{1,n})$ within the extremum points, Lemma 2.6 in [96] and Lemma 3.13 together with Theorem 3.10 result in

$$|\frac{d\lambda}{d\varphi}(c, \tilde{\varphi})| \leq |\frac{d\lambda}{d\varphi}(c_{1,n}, \tilde{\varphi}_{1,n})|, \tag{3.43}$$

where $\tilde{\varphi}$ and $\tilde{\varphi}_{1,n}$ define points of equal Rayleigh quotients on $S(c)$ and $S(c_{1,n})$, respectively.

Now $f(\varphi) = \lambda(c_{1,n}, \varphi)$ and $g(\varphi) = \lambda(c, \varphi)$ are monotone decreasing, differentiable positive functions. Equation (3.43) simply says that in all arguments α, β having the same value $f(\alpha) = g(\beta)$, the (negative) derivatives fulfill

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

Hence, with $f(0) = g(0)$ it holds

$$f(a - \xi) \leq g(b - \xi),$$

where we next identify ξ with the smaller angle $\xi = \varphi^* + \varphi_\gamma(c)$. We conclude for the even larger angle $\varphi^*_{1,n} + \varphi_\gamma(c_{1,n})$ that

$$\lambda(c_{1,n}, \varphi^*_{1,n} + \varphi_\gamma(c_{1,n})) \leq \lambda(c, \varphi^* + \varphi_\gamma(c)),$$

which proves the first proposition. The Rayleigh quotient $\lambda' = \lambda^-_{1,n}(\lambda, \gamma)$ is a consequence of Theorem 3.14 for $i = 1$ and $j = n$ and the best preconditioning belonging to "–" in (3.37).

To show (2), we proceed analogously as in the proof of Theorem 1.1 in [96] but compare the curves of suprema points of c and $c_{1,n} \in L(\lambda)$. Then for the opening and shrinking angles it holds that

$$\varphi_\gamma(c) \leq \varphi_\gamma(c_{1,n}) \quad \text{and} \quad \psi_\gamma(c) \leq \psi_\gamma(c_{1,n}).$$

Instead of (3.1) in [96] one has

$$\left| \frac{d\bar{\lambda}}{d\varphi}(c, \varphi^*) \right| \leq \left| \frac{d\bar{\lambda}}{d\varphi}(c_{1,n}, \varphi^*_{1,n}) \right|.$$

Comparing the decrease of $\bar{\lambda}(c, \varphi)$ and $\bar{\lambda}(c_{1,n}, \varphi)$ shows that

$$\bar{\lambda}(c, \varphi_1(c) - \psi_\gamma(c)) \geq \bar{\lambda}(c_{1,n}, \varphi_1(c_{1,n}) - \psi_\gamma(c)),$$

which proves the second proposition since

$$\bar{\lambda}(c, \varphi_\gamma(c)) > \bar{\lambda}(c_{1,n}, \varphi_\gamma(c_{1,n})).$$

The Rayleigh quotient $\lambda^+_{1,n}$ results from applying the mini-dimensional analysis to the 2D space $\mathrm{span}\{e_1, e_n\}$, see Section 3.3.

Finally, for the proof of (3) we proceed similarly to (1) but compare $c,\ c_{i,i+1} \in L(\lambda)$. Let us remark that the Rayleigh quotient $\lambda^-_{i,i+1}(\lambda, \gamma)$ provides only an estimate from above since by the construction all infima are forced to be located in $\mathrm{span}\{e_i, e_{i+1}\}$. But as a matter of fact, the bifurcation of the infimum curve by introducing components to e_1 hastens the decrease of the Rayleigh quotient. □

Let us illustrate the convergence estimates derived in Theorem 3.15 on the basis of the example used in Section 2.4, i.e. the eigenvalue problem for Δ_h in $\mathbb{R}^2$ with eigenvalues given by (2.30). In Figure 3.5 the quotients

$$\Phi^\pm_{i,j}(\lambda, \gamma) := \frac{\lambda^\pm_{i,j}(\lambda, \gamma) - \lambda_i}{\lambda - \lambda_i}, \tag{3.44}$$

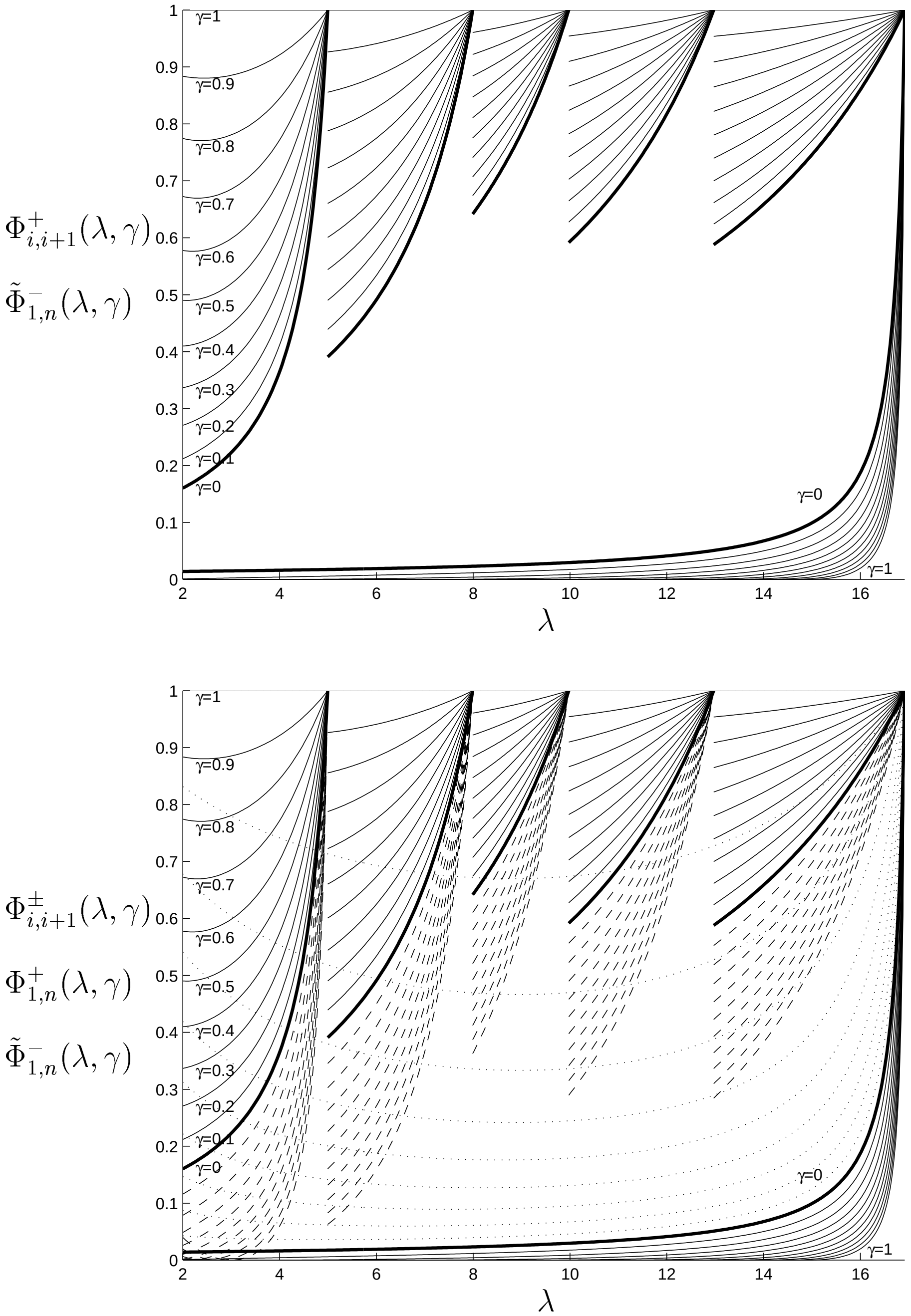


Figure 3.5: $\boxed{\substack{a\\b}}$ *PINVIT convergence estimates. (a) Poorest ($\Phi^{+}_{i,i+1}$) and fastest ($\tilde{\Phi}^{-}_{1,n}$) convergence, $\gamma = 0, 0.1, \ldots, 1$. (b) Supplementing the remaining combinations: $\Phi^{-}_{i,i+1}$ by dashed lines and $\Phi^{+}_{1,n}$ by dotted lines.*

$\lambda_{i,j}^{\pm}(\lambda,\gamma)$ defined in Theorem 3.14, are drawn for $\lambda \in [2,17]$. First, in the upper part of Figure 3.5 the estimates $\Phi_{i,i+1}^{+}(\lambda,\gamma)$, as already displayed in Figure 2.2, are supplemented by the bounds $\Phi_{1,n}^{-}(\lambda,\gamma)$ reflecting the fastest decrease of the Rayleigh quotient. We remark that the assumption $e_1 \notin C_\gamma(c)$, as made in Theorem 3.15, is only made to avoid tiresome case distinctions. Whenever for some $\lambda^* \in [\lambda_1,\lambda_n]$ it holds that $\lambda_{1,n}^{-}(\lambda^*,\gamma^*) = \lambda_1$, then $e_1 \in C_\gamma(c)$ for any γ larger than γ^*. Hence, what actually is drawn in Figure 3.5 is

$$\tilde{\Phi}_{1,n}^{-}(\lambda,\gamma) := \min_{\tilde{\gamma}\leq\gamma} \Phi_{1,n}^{-}(\lambda,\tilde{\gamma}). \tag{3.45}$$

Finally, in the lower part of Figure 3.5 the remaining combinations 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 dashed lines.

This numerical example shows a wide corridor between best and poorest convergence. While the estimates on poorest convergence in $[\lambda_i,\lambda_{i+1}]$ do not depend on the largest eigenvalue λ_n (we exclude the trivial case $i+1=n$), the estimates $\Phi_{1,n}^{\pm}$ on the faster convergence do so. Hence, whenever λ_n increases, the corridor between poorest and best convergence widens, allowing increasingly faster convergence. In $\operatorname{span}\{e_1,e_n\}$ we can also determine the particular λ^*, below which PINVIT(1) is capable of converging to e_1 in only one step. Condition (3.2) in $\operatorname{span}\{e_1,e_n\}$ leads to

$$\lambda^* = \lambda_n \left(1 + \frac{\lambda_1}{\gamma^2(\lambda_n-\lambda_1)}\right)^{-1}. \tag{3.46}$$

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 (3.46) and solving for γ results in

$$\gamma^* = \frac{\lambda_1}{\lambda_n - \lambda_1}.$$

Therefore, a large λ_n, as $\lambda_n \simeq h^{-2}$ for Δ_h, makes one-step convergence under weak conditions possible, cf. the curves $\tilde{\Phi}_{1,n}^{-}$ in Figure 3.5. The bold curves in Figure 3.5 are associated with $\gamma = 0$ or inverse iteration. Not surprisingly, PINVIT may converge faster than INVIT as the most favorable choice of the preconditioner hastens convergence compared to INVIT. The upper bold curves correspond to $c_{i,i+1}$, $i = 1,\dots,4$, while the lower bold curve reflects the fastest possible convergence of INVIT in $c_{1,n}$.

Finally, Figure 3.6 is the pendant of Figure 1.2 shown in Section 1.4 containing the model analysis of INVIT(1). The bold curves ($\gamma = 0$) in Figure 3.6 are identical with the bounds $B(\lambda_1,\lambda_n,\lambda)$ and $B(\lambda_i,\lambda_{i+1},\lambda)$, for $i = 1,\dots,4$, as drawn in Figure 1.2. The lower curves $\tilde{\lambda}_{1,n}^{-}(\lambda,\gamma)$ for $\gamma = 0, 0.1, \dots, 1$ are defined in a similar way as $\tilde{\Phi}_{1,n}(\lambda,\gamma)$ in (3.45).

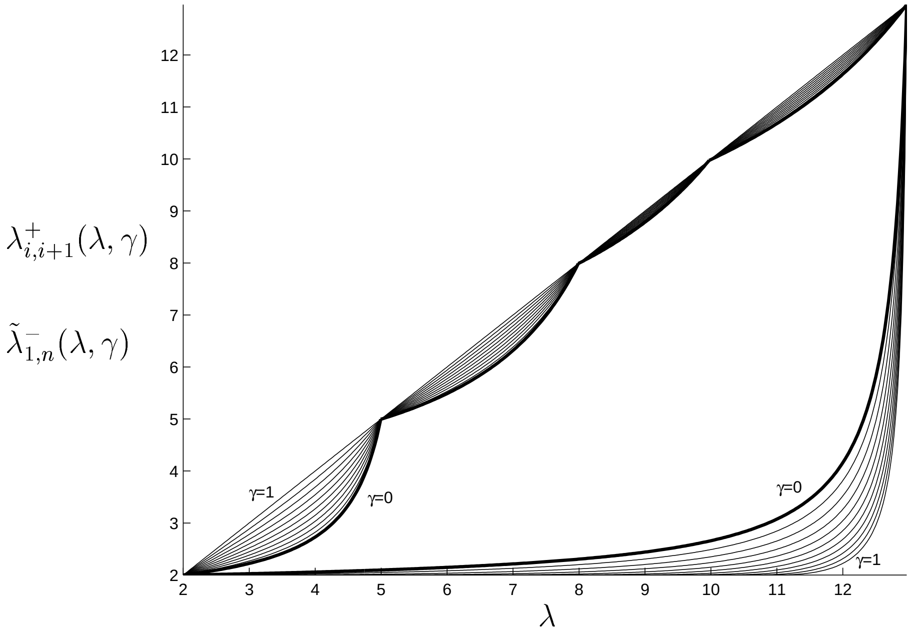


Figure 3.6: *Plot of* $\lambda^+_{i,i+1}(\lambda,\gamma)$ *(upper curves) and* $\tilde\lambda^-_{1,n}(\lambda,\gamma)$ *(lower curves) for* $\gamma = 0, 0.1, \ldots, 1$.

3.5 Critical conclusion

- Sharp convergence estimates for PINVIT(1) have been derived. Most of these estimates are sharp in λ, γ, λ_i, λ_{i+1} or λ, γ, λ_1, λ_n, respectively.
- These estimates substantiate that PINVIT(1) may converge much more rapidly than suggested by Theorem 2.8 on the poorest PINVIT convergence. PINVIT may even converge faster than INVIT (the case of exact preconditioning). Under the (weak) condition of Lemma 3.2 *one-step convergence to an eigenpair is possible*.
- The main drawback of the convergence estimates presented so far is their complicated dependence on λ, γ and λ_i, λ_{i+1} or λ_1, λ_n, respectively.
- The bounds of Theorems 2.8 and 3.14 do not allow a simple recursive representation for estimating multiple-step convergence.

Some remedy overcoming the listed disadvantages will be given in Chapter 4.

4. Concise convergence estimates

In Chapters 2 and 3 sharp upper and lower bounds for the Rayleigh quotient of the PINVIT(1) iterates have been derived. These estimates, as given in Theorems 2.8 and 3.15, suffer from the complexity of Equations (2.27) and (3.37) with their complicated dependence of $\lambda_{i,j}^{\pm}$ on λ_i, λ_j, λ and γ. This makes it difficult to judge on the quality of the convergence estimates and impedes any comparison with the convergence factors of other simplified or improved (preconditioned) eigensolvers. These circumstances establish the need to derive simplified estimates without sacrificing too much sharpness.

Here, we derive easy-to-read convergence estimates for the Rayleigh quotient in terms of the $\Delta_{p,q}$ factors as introduced in the model analysis of INVIT(1) in Section 1.4. The resulting bounds are sharp in λ_i, λ_{i+1} and γ, but only asymptotically sharp in λ. They turn into the previously derived estimates as λ tends to λ_i. The (minor) drawback of losing (non-asymptotic) sharpness in λ is compensated by having easy-to-use convergence factors which allow a recursive representation, too.

The results presented in this chapter originate from a joint work with Andrew Knyazev, University of Colorado at Denver; see also [73]. In the following these new bounds will be gained in two distinct ways:

1. In Section 4.1 we use the results of the mini-dimensional analysis of Section 3.3 to find an advantageous compact representation of $\Delta_{p,q}(\lambda')/\Delta_{p,q}(\lambda)$ in terms of those geometric quantities which define the geometry of PINVIT in $\mathrm{span}\{x_p, x_q\}$. This analysis does not provide an independent proof of PINVIT convergence, since it makes use of the complete PINVIT theory including the mini-dimensional analysis.

2. The second approach is built upon certain matrix functions designed to represent the extremum points of $\lambda(\cdot)$ on the ball E_γ, i.e. we make use of the representation

$$F(\alpha, x) = \beta(\alpha I + A)^{-1}x \tag{4.1}$$

   of points of absolute extrema, cf. Theorem 3.9 and Theorem 4.8 in [95]. This proof technique may have the potential to make a new succinct convergence proof possible for PINVIT as well as for improved conjugate-gradient like preconditioned eigensolvers as

PINVIT(3,s), or LOBPCG. The idea of this promising technique goes back to Knyazev [67, 68]. But at the moment the explicit form of α as a function of λ_i, λ_j, λ and γ in 2D, see Equation (3.18), is needed. Presently, no simple formula is known unveiling this dependence in the general case. Thus the present analysis is just an alternative form of the mini-dimensional analysis given in [95], Section 5. The further analysis is currently under investigation as a joint work with A. Knyazev.

4.1 Reformulation of the mini-dimensional analysis

By attacking the mini-dimensional PINVIT analysis (see Theorem 5.1 in [95]), Theorem 4.1 provides a succinct form of the PINVIT estimates; see also [73]. A recursive formulation is given by Equation (4.4).

Theorem 4.1. *Let a nonzero vector $x^{(0)} \in \mathbb{R}^n$ be given and let $(x^{(j)}, \lambda^{(j)})$ be the sequence of PINVIT iterates for $j = 0, 1, 2, \ldots$. The preconditioner is assumed to satisfy (2.2) for $\gamma \in [0, 1)$. If $\lambda^{(j)} = \lambda(x^{(j)}) \in [\lambda_k, \lambda_{k+1})$ then either $\lambda^{(j+1)} < \lambda_k$ or*

$$\Delta_{k,k+1}(\lambda^{(j+1)}) \leq \left(\gamma + (1-\gamma)\frac{\lambda_k}{\lambda_{k+1}}\right)^2 \Delta_{k,k+1}(\lambda^{(j)}), \tag{4.2}$$

where

$$\Delta_{k,k+1}(\lambda) = \frac{\lambda - \lambda_k}{\lambda_{k+1} - \lambda}. \tag{4.3}$$

Whenever $\lambda^{(j)} < \lambda_2$, then for $m = 1, 2, \ldots$ the recursive estimate

$$\Delta_{1,2}(\lambda^{(j+m)}) \leq \left(\gamma + (1-\gamma)\frac{\lambda_1}{\lambda_2}\right)^{2m} \Delta_{1,2}(\lambda^{(j)}) \tag{4.4}$$

holds.

The estimate (4.2) is sharp in λ_k, λ_{k+1} and γ. It is asymptotically sharp in λ and turns into a sharp estimate (cf. Theorem 2.8) as $\lambda \to \lambda_k$.

Proof. We make use of the notation concerning the geometric quantities used in the mini-dimensional analysis in Theorem 5.1 in [95]; for the definition and meaning of ξ, η, x, y, r and l see the very place. Without loss of generality let $k = 1$. The Rayleigh quotient $\lambda_{1,2}(\lambda, \gamma)$ within the supremum on E_γ reads

$$\lambda_{1,2}(\lambda, \gamma) = \frac{\eta^2 + \xi^2}{\eta^2/\lambda_1 + \xi^2/\lambda_2}, \tag{4.5}$$

where (η, ξ) are the coordinates of the supremum point, see Equation (5.6) in [95]. Then the ratio

$$\Delta_{1,2}(\lambda_{1,2}) = \frac{\lambda_{1,2} - \lambda_1}{\lambda_2 - \lambda_{1,2}}$$

by inserting (4.5) as well as

$$(\eta,\xi) = (\sqrt{l^2-\xi^2},\ \frac{xl^2+ryl}{x^2+y^2}), \tag{4.6}$$

is transformed to

$$\begin{aligned}\frac{\lambda_{1,2}-\lambda_1}{\lambda_2-\lambda_{1,2}} &= \frac{\xi^2\lambda_1}{\eta^2\lambda_2}\\ &= \frac{\lambda_1}{\lambda_2}\,\frac{(xl+ry)^2}{(x^2+y^2)^2-(xl+ry)^2}.\end{aligned}$$

We have

$$\sqrt{(x^2+y^2)^2-(xl+ry)^2} = \pm(yl-xr),$$

where $yl-xr$ is the positive root because of $y>r$.

For the quotient of c_1 and c_2, see Equation (3.17), it holds

$$\left(\frac{c_1}{c_2}\right)^2 = \frac{\lambda_1(\lambda_2-\lambda)}{\lambda_2(\lambda-\lambda_1)},$$

so that

$$\frac{\lambda_2-\lambda}{\lambda-\lambda_1} = \frac{\lambda_2 c_1^2}{\lambda_1 c_2^2} = \frac{y^2\lambda_1}{x^2\lambda_2}.$$

Then the convergence factor σ, defined by

$$\frac{\lambda_{1,2}-\lambda_1}{\lambda_2-\lambda_{1,2}}\cdot\frac{\lambda_2-\lambda}{\lambda-\lambda_1} = \frac{\lambda_1^2y^2}{\lambda_2^2x^2}\,\frac{(xl+ry)^2}{(yl-rx)^2} =: \sigma^2, \tag{4.7}$$

reads

$$\sigma = \frac{\lambda_1 y(xl+ry)}{\lambda_2 x(yl-rx)} = \frac{\lambda_1}{\lambda_2}\cdot\frac{1+\frac{yr}{xl}}{1-\frac{xr}{yl}} > 0. \tag{4.8}$$

Direct computation shows that

$$\frac{yr}{xl} = \gamma(\lambda_2-\lambda)\left(\frac{\lambda_2}{\lambda_1}\right)^{1/2} z^{-1/2}$$

and

$$\frac{xr}{yl} = \gamma(\lambda-\lambda_1)\left(\frac{\lambda_1}{\lambda_2}\right)^{1/2} z^{-1/2}$$

with $z := \gamma^2(\lambda_1-\lambda)(\lambda_2-\lambda)+\lambda(\lambda_1+\lambda_2-\lambda) > 0$. Hence

$$\sigma[\lambda] = \frac{\left(\frac{\lambda_1}{\lambda_2}\right)^{1/2} z^{1/2}+\gamma(\lambda_2-\lambda)}{\left(\frac{\lambda_2}{\lambda_1}\right)^{1/2} z^{1/2}-\gamma(\lambda-\lambda_1)}. \tag{4.9}$$

For $\lambda = \lambda_1$, we have

$$\sigma[\lambda_1] = \gamma + (1-\gamma)\frac{\lambda_1}{\lambda_2},$$

or

$$\sigma[\lambda_1] = \frac{\lambda_1}{\lambda_2} + \gamma(1 - \frac{\lambda_1}{\lambda_2}).$$

To complete the proof we show that

$$\sigma'[\lambda] < 0,$$

which is equivalent to

$$\gamma\sqrt{\lambda_1\lambda_2}(\lambda_2-\lambda_1) < (\lambda_2-\lambda_1)z^{1/2} + (\frac{d}{d\lambda}z^{1/2})\left\{\lambda_2(\lambda_2-\lambda)+\lambda_1(\lambda-\lambda_1)\right\}.$$

The last inequality is true, since for its square (by inserting z as well as $\frac{d}{d\lambda}z^{1/2}$ and subsequent factorization) it holds the true inequality

$$(1-\gamma^2)(\lambda_2-\lambda_1)^2(\lambda_1+\lambda_2-\lambda)^2\left[(1+\gamma)\lambda_1+(1-\gamma)\lambda_2\right]\left[(1-\gamma)\lambda_1+(1+\gamma)\lambda_2\right] > 0.$$

Sharpness of (4.2) in λ_1, λ_2 and γ is a consequence of the construction of (4.9) and Theorem 1.1 in [96]. The asymptotic sharpness for $\lambda \to \lambda_1$ follows from (4.9), too. □

4.2 A matrix function approach

The idea of this approach is to restrict the PINVIT scheme only on those preconditioners responsible for the best and poorest convergence. By Theorem 4.8 in [95] (case of suprema) and Theorem 3.9 (case of infima) those points of absolute extrema can be represented in the form of (4.1), i.e. inverse iteration with a positive shift (suprema) or negative shift (infima), respectively. In Section 4.2.1 we first describe a general setup for iterative eigensolvers induced by certain matrix functions, which damp out the invariant subspace of A belonging to the eigenvalues $\lambda_2, \ldots, \lambda_n$. These results will be used later (in Section 4.2.2) within the PINVIT setup.

4.2.1 An abstract convergence estimate

Only for this section assume $A \in \mathbb{R}^{n\times n}$ to be a symmetric and positive definite matrix with the eigenvalues $\lambda_1 \le \ldots \le \lambda_n$ of arbitrary multiplicity. In the preparatory Lemma 4.2 we formulate an abstract condition showing that the Rayleigh quotient is increased, if some relative damping is applied to its argument in a way that the spectral components belonging to the smaller eigenvalues are decreased.

Lemma 4.2. *Suppose a nonzero $x \in \mathbb{R}^n$ with $\lambda_m \leq \lambda(x) < \lambda_{m+1}$ to be given. Its expansion in normed eigenvectors x_i of A is written as $x = \sum_{i=1}^n c_i x_i$. If for the components a_i of $a \in \mathbb{R}^n$ it holds that*

$$\max\{|a_1|, \ldots, |a_m|\} \leq \min\{|a_{m+1}|, \ldots, |a_n|\},$$

then $y := \sum_{i=1}^n a_i c_i x_i$ satisfies $\lambda(x) \leq \lambda(y)$.

Proof. Let $|a_k| = \min\{|a_{m+1}|, \ldots, |a_n|\}$. If $a_k = 0$, then $a_1 = \ldots = a_m = 0$ and then $\lambda(y) \geq \lambda_{m+1}$. Next assume $|a_k| > 0$ so that

$$\lambda(y) = \frac{\sum_{i \leq m} (a_i/a_k)^2 c_i^2 \lambda_i + c_k^2 \lambda_k + \sum_{i \geq m+1,\, i \neq k} (a_i/a_k)^2 c_i^2 \lambda_i}{\sum_{i \leq m} (a_i/a_k)^2 c_i^2 + c_k^2 + \sum_{i \geq m+1,\, i \neq k} (a_i/a_k)^2 c_i^2}.$$

A direct computation shows that any decrease of components $i \leq m$ by $(a_i/a_k)^2 \leq 1$ or any increase of components $i \geq m+1$ by $(a_i/a_k)^2 \geq 1$ results in an increased Rayleigh quotient which proves the assertion. □

Lemma 4.2 proves Lemma 2.3.2 in [68] reproduced as a corollary.

Corollary 4.3. *Let $F = F(A)$ be a real matrix function [45] of A and assume $x \in \mathbb{R}^n$ with $\lambda_m \leq \lambda(x) < \lambda_{m+1}$ to be given. If*

$$|F(\lambda_i)| \leq 1, \quad i = 1, \ldots, m, \quad \textit{and} \quad |F(\lambda_i)| \geq 1, \quad i = m+1, \ldots, n,$$

then $\lambda(x) \leq \lambda(F(A)x)$.

Now, consider the linear iterative scheme

$$x' = Fx \tag{4.10}$$

associated with the matrix function $F = F(A)$, which maps a given iterate x to the next iterate x'. Theorem 4.4 formulates a condition under which (4.10) can serve as an eigensolver. This theorem is reproduced from Theorem 2.3.1 in [68] (in Russian); it also appeared (without a proof) as an English translation in [69].

Theorem 4.4. *For any nonzero $x \in \mathbb{R}^n$ with $\lambda(x) \in (\lambda_1, \lambda_2)$ and under the assumption that $F(\lambda_1) \neq 0$ and*

$$\sigma := \max_{i>1} \left| \frac{F(\lambda_i)}{F(\lambda_1)} \right| < 1$$

we have

$$\Delta\lambda(x') \leq \sigma^2 \Delta\lambda(x), \tag{4.11}$$

where

$$\Delta\lambda(y) := \frac{\lambda(y) - \lambda_1}{\lambda_2 - \lambda(y)}.$$

Proof. Let us define $\hat{F}$ as follows

$$\begin{aligned}\hat{F}(\lambda_1) &= F(\lambda_1)\\ \hat{F}(\lambda_i) &= \max_{j>1}|F(\lambda_j)| = \sigma F(\lambda_1) \quad \text{for} \quad 1 < i \leq n,\end{aligned}$$

and $\hat{x} = \hat{F}(A)x$. From Corollary 4.3 we obtain

$$\lambda(x') \leq \lambda(\hat{x}) \leq \lambda(x). \tag{4.12}$$

We define the 2-dimensional space $H^{[2]} := \mathrm{span}\{x, \hat{x}\} = \mathrm{span}\{x, \hat{x}, x_1\}$, where (x_1, λ_1) denotes the smallest eigenpair of A. The last equality holds since $\hat{F}$ treats all spectral components of x different to x_1 in the same way.

Let $A^{[2]} = P^{[2]}A$ be the projection of A to $H^{[2]}$ and let Q be the orthoprojector on $\mathrm{span}\{x_1\}$. Then Qx and $(I-Q)x$ are the eigenvectors of $A^{[2]}$, since Qx as a multiple of x_1 minimizes the Rayleigh quotient and $(Qx, (I-Q)x) = 0$. The eigenvalues of $A^{[2]}$ are $\lambda(Qx) = \lambda_1$ and the Rayleigh quotient of $(I-Q)x$ with respect to A or $A^{[2]}$.

Direct computation using

$$\begin{aligned}x &= Qx + (I-Q)x\\ \hat{x} &= \hat{F}x = F(\lambda_1)(Qx + \sigma((I-Q)x))\end{aligned}$$

together with $((I-Q)x, AQx) = 0$ shows that

$$\frac{\lambda(\hat{x}) - \lambda_1}{\lambda((I-Q)x) - \lambda(\hat{x})} = \sigma^2 \frac{\lambda(x) - \lambda_1}{\lambda((I-Q)x) - \lambda(x)}. \tag{4.13}$$

Recognizing that the quotient on the left-hand side of (4.13) is a monotone increasing function in $\lambda(\hat{x})$ together with (4.12) results in

$$\frac{\lambda(x') - \lambda_1}{\lambda((I-Q)x) - \lambda(x')} \cdot \frac{\lambda((I-Q)x) - \lambda(x)}{\lambda(x) - \lambda_1} \leq \sigma^2.$$

Finally, note that the left-hand side is an increasing function in $\lambda((I-Q)x)$. The Courant-Fischer principle implies $\lambda((I-Q)x) > \lambda_2$ which establishes Equation (4.11). □

4.2.2 Several estimates on extremal convergence

The aim of this section is to apply Theorem 4.4 in order to derive convergence estimates for preconditioned inverse iteration. The central idea is to consider the curve $F(\alpha, A)x = \beta(\alpha I + A)^{-1}x$ through the points of absolute extrema as derived in Theorem 4.8 in [95] concerning points of suprema and in Theorem 3.9 of this work concerning points of infima.

We first observe that points of extrema can be represented by applying the operator $\beta(\alpha I + A)^{-1}$ to the given iterate x. The latter operator depends in a complex and nonlinear way on the iterate x, as α and β are function of x and γ; cf. the discussion in Section 3.1.4. Obviously, the scaling constant β is meaningless as it does not influence the Rayleigh quotient of $F(\alpha, A)x$. In the following, we therefore set $\beta = 1$. Unfortunately, the shift parameter α causes some trouble because of its *implicit* definition by (3.16). Only within 2D invariant subspaces of A, Equation (3.18) gives an *explicit* formula for α. In $\operatorname{span}\{x_i, x_j\}$, which is the invariant subspace to the eigenvalues λ_i and λ_j, $\lambda_i < \lambda_j$, it holds

$$\alpha^{\pm} = \frac{\gamma\sqrt{\lambda_i\lambda_j}}{\lambda(1-\gamma^2)}\left(\gamma\sqrt{\lambda_i\lambda_j} \pm \sqrt{(1-\gamma^2)(\lambda_j-\lambda)(\lambda-\lambda_i)+\lambda_i\lambda_j}\right) \tag{4.14}$$

where $\alpha^+ > 0$ ($\alpha^- < 0$) belongs to a supremum (infimum) point. The choice $\alpha = 0$ corresponds to inverse iteration, i.e. $\gamma = 0$. Then best and poorest convergence coincide in $\lambda A^{-1}x$, being the only element in $E_0(x)$.

Consequently, any further analysis using Equation (4.14) is restricted to 2D invariant subspaces of A. But even an application to such 2D spaces is worthwhile as the mini-dimensional analysis, cf. Section 3.3, is performed in a 2D invariant subspace of A. In the following such an analysis will result in a compact representation of the PINVIT convergence estimates, both for the best and the poorest decrease of the Rayleigh quotient. We note that the resulting concise convergence estimates presuppose the justification for the mini-dimensional analysis, which is given by the "angle analysis on $L(\lambda)$" ([96, Sections 2 and 3]) and Section 3.2 in this work.

As already mentioned in the introduction to Chapter 4, the following analysis, which is based on Equation (4.14), only reflects the current state of research. We hope that a more general representation of α in the $\mathbb{R}^n$ can serve to simplify the PINVIT convergence analysis considerably.

In the following (preparatory) Lemma 4.5 we state some monotonicity of $\alpha^{\pm}$, which will be used in Theorem 4.6.

Lemma 4.5. *The shifts $\alpha^{\pm}$ are strictly monotone functions of $\lambda \in [\lambda_i, \lambda_j]$. It holds that*

$$\frac{\partial\alpha^+}{\partial\lambda} < 0 \quad \textit{and} \quad \frac{\partial\alpha^-}{\partial\lambda} > 0.$$

Proof. We obtain

$$\begin{aligned}\frac{\partial\alpha^{\pm}}{\partial\lambda}\,\frac{\lambda(1-\gamma)}{\gamma\sqrt{\lambda_i\lambda_j}} &= -\frac{\gamma\sqrt{\lambda_i\lambda_j}}{\lambda} \mp \frac{1}{\lambda}\left((1-\gamma^2)(\lambda_j-\lambda)(\lambda-\lambda_i)+\lambda_i\lambda_j\right)^{1/2}\\ &\quad\pm \frac{(1-\gamma^2)(\lambda_i+\lambda_j-2\lambda)}{2\left((1-\gamma^2)(\lambda_j-\lambda)(\lambda-\lambda_i)+\lambda_i\lambda_j\right)^{1/2}}.\end{aligned} \tag{4.15}$$

To show $(\partial\alpha^+/\partial\lambda) < 0$ we have to prove that the right-hand side of (4.15), signs corresponding to α^+, is negative. After simplification we obtain

$$0 \le \gamma\sqrt{\lambda_i\lambda_j}\left((1-\gamma^2)(\lambda_j-\lambda)(\lambda-\lambda_i)+\lambda_i\lambda_j\right)^{1/2} + (1-\gamma^2)\frac{\lambda(\lambda_i+\lambda_j)}{2} + \gamma^2\lambda_i\lambda_j,$$

which is clearly true, since all summands are positive.

Finally, $(\partial\alpha^-/\partial\lambda) > 0$ is equivalent to

$$(1-\gamma^2)\frac{\lambda(\lambda_i+\lambda_j)}{2} + \gamma^2\lambda_i\lambda_j > \gamma\sqrt{\lambda_i\lambda_j}\left((1-\gamma^2)(\lambda_j-\lambda)(\lambda-\lambda_i)+\lambda_i\lambda_j\right)^{1/2}.$$

Both sides of the last inequality are positive. Hence, after squaring and subsequent factorization we obtain the following equivalent inequality

$$0 < \frac{1}{4}\lambda^2(1-\gamma^2)\left((\lambda_j-\lambda_i)\gamma+\lambda_i+\lambda_j\right)\left(\lambda_i(1-\gamma)+\lambda_j(1+\gamma)\right).$$

All factors on the right-hand side are positive, which completes the proof. □

Theorem 4.6 makes available concise formula for the convergence factors concerning the poorest and best convergence of PINVIT depending on the choice of the preconditioner and the iteration vector x. The convergence factor σ_1 on the poorest convergence is the same as the one which has been derived by a different technique in Theorem 4.1.

Theorem 4.6. *Let $x^{(0)} \in \mathbb{R}^n$ and denote the PINVIT iterates by $(x^{(j)}, \lambda^{(j)})$. The preconditioner for some $\gamma \in [0,1)$ obeys the quality condition (2.2).*

If $\lambda^{(j)} \in [\lambda_k, \lambda_{k+1})$ then either $\lambda^{(j+1)} < \lambda_k$ or

$$\Delta_{k,k+1}(\lambda^{(j+1)}) \le \sigma_1^2\Delta_{k,k+1}(\lambda^{(j)}), \tag{4.16}$$

with

$$\sigma_1 := \gamma + (1-\gamma)\frac{\lambda_k}{\lambda_{k+1}}. \tag{4.17}$$

For the best choice of the preconditioner in $\mathcal{B}_\gamma$, by Equation (2.3), and for the most advantageous selection of $x \in L(\lambda^{(j)})$ the convergence estimate reads

$$\Delta_{1,n}(\lambda^{(j+1)}) \le \sigma_7^2\Delta_{1,n}(\lambda^{(j)}), \tag{4.18}$$

with

$$\sigma_7 := \frac{\lambda_1}{\lambda_n+\gamma(\lambda_n-\lambda_1)}. \tag{4.19}$$

We denote the latter convergence factor by σ_7 by reason of some systematics to be introduced later in Table 4.1.

Proof. Just apply Theorem 4.4 to

$$x' = Fx = \beta(\alpha I + A)^{-1}x,$$

representing the unique supremum point of $\lambda(\cdot)$ on $E_\gamma(x)$. The analysis in [96] provides the justification to restrict the further analysis to the 2D space $\operatorname{span}\{x_k, x_{k+1}\}$. Then

$$\sigma = \left|\frac{F(\lambda_{k+1})}{F(\lambda_k)}\right| = \frac{\alpha+\lambda_k}{\alpha+\lambda_{k+1}}.$$

Next observe that $\sigma[\alpha]$ is strictly monotone increasing in α. Furthermore, $\alpha = \alpha^+$ by Equation (4.14) is strictly monotone decreasing in λ by Lemma 4.5 so that its maximum is taken in

$$\alpha^+[\lambda_k] = \frac{\gamma}{1-\gamma}\lambda_{k+1}.$$

Finally, we get the asymptotically sharp estimate

$$\frac{F(\lambda_{k+1})}{F(\lambda_k)} \leq \frac{\alpha^+[\lambda_k] + \lambda_k}{\alpha^+[\lambda_k] + \lambda_{k+1}} = \gamma + (1-\gamma)\frac{\lambda_k}{\lambda_{k+1}} = \sigma_1. \tag{4.20}$$

To show the convergence estimate on the best convergence, Theorem 3.15 prescribes the mini-dimensional analysis in $\mathrm{span}\{x_1, x_n\}$. For an infimum point we have to deal with negative α^-, with $-\lambda_1 < \alpha^- < 0$, i.e.

$$\sigma = \left|\frac{F(\lambda_n)}{F(\lambda_1)}\right| = \frac{\alpha^- + \lambda_1}{\alpha^- + \lambda_n}.$$

Since by Lemma 4.5 the function $\alpha^-[\lambda]$ is strictly monotone increasing, its maximum reads

$$\alpha^-[\lambda_n] = -\frac{\gamma\lambda_1}{1+\gamma}.$$

Hence,

$$\frac{F(\lambda_n)}{F(\lambda_1)} \leq \frac{\alpha^-[\lambda_n] + \lambda_1}{\alpha^-[\lambda_n] + \lambda_n} = \frac{\lambda_1}{\lambda_n + \gamma(\lambda_n - \lambda_1)} = \sigma_7.$$

□

Theorem 4.6 only contains the most important convergence factors. The remaining combinations of the poorest and best choice of the preconditioner $B^{-1} \in \mathcal{B}_\gamma$ and the iteration vector $x \in L(\lambda)$ are listed in Table 4.1. The λ-column of Table 4.1 reflects the choice of the minimum or maximum of $\alpha^{\pm}[\lambda]$. The '$+$' ('$-$') symbol indicates a choice maximizing (minimizing) the Rayleigh quotient. Remember that this additional degree of freedom (in comparison with the PINVIT convergence Theorems 2.8 and 3.15) is the price we have to pay for attaining the succinct form of the convergence factors σ_i.

Let us mention that $\sigma_5, \ldots, \sigma_8$ are *decreasing* functions of γ since expanding the ball $E_\gamma(x)$ decreases the smallest attainable Rayleigh quotient. The factors σ_6 and σ_8 may take the value 0 which reflects the fact that an eigenvector to λ_1 is contained in the ball E_γ, which makes one-step convergence possible. Corollary 4.7 compiles some relations between the σ_i. It is proved by simple algebraic manipulations.

Corollary 4.7. *For the convergence factors σ_i it holds:*

1. $\sigma_1 \geq \sigma_2 \geq \sigma_6 \geq \sigma_8,$

i	B	x	λ	σ_i
1	$-$	$-$	$-$	$\gamma + (1-\gamma)\dfrac{\lambda_1}{\lambda_2}$
2	$-$	$-$	$+$	$\dfrac{\lambda_1}{\lambda_2 - \gamma(\lambda_2 - \lambda_1)}$
3	$-$	$+$	$-$	$\gamma + (1-\gamma)\dfrac{\lambda_1}{\lambda_n}$
4	$-$	$+$	$+$	$\dfrac{\lambda_1}{\lambda_n - \gamma(\lambda_n - \lambda_1)}$
5	$+$	$-$	$-$	$\dfrac{\lambda_1}{\lambda_2 + \gamma(\lambda_2 - \lambda_1)}$
6	$+$	$-$	$+$	$\max(0, \dfrac{\lambda_1}{\lambda_2} - \gamma(1 - \dfrac{\lambda_1}{\lambda_2}))$
7	$+$	$+$	$-$	$\dfrac{\lambda_1}{\lambda_n + \gamma(\lambda_n - \lambda_1)}$
8	$+$	$+$	$+$	$\max(0, \dfrac{\lambda_1}{\lambda_n} - \gamma(1 - \dfrac{\lambda_1}{\lambda_n}))$

Table 4.1: *PINVIT convergence factors σ_i.*

2. $\sigma_1 \geq \sigma_3 \geq \sigma_7 \geq \sigma_8$*,*

3. $\sigma_2 \geq \sigma_4$ *and* $\sigma_3 \geq \sigma_4$*,*

4. $\sigma_5 \geq \sigma_6$ *and* $\sigma_5 \geq \sigma_7$*,*

or schematically,

$$\begin{tikzcd}
 & \sigma_2 \arrow[dr, phantom, sloped, "\geq"] \arrow[rrr, phantom, "\geq"] & & & \sigma_6 \arrow[dr, phantom, sloped, "\geq"] & \\
\sigma_1 \arrow[ur, phantom, sloped, "\geq"] \arrow[dr, phantom, sloped, "\geq"] & & \sigma_4 \arrow[r, phantom, "\vdots"] & \sigma_5 \arrow[ur, phantom, sloped, "\geq"] \arrow[dr, phantom, sloped, "\geq"] & & \sigma_8 \, . \\
 & \sigma_3 \arrow[ur, phantom, sloped, "\geq"] \arrow[rrr, phantom, "\geq"] & & & \sigma_7 \arrow[ur, phantom, sloped, "\geq"] &
\end{tikzcd}$$

In the next corollary we show that one-step convergence to an eigenvector belonging to the smallest eigenvalue is possible if $\gamma \geq \lambda_1/(\lambda_n - \lambda_1)$; compare also the discussion on the fastest convergence of the PINVIT scheme in the introduction of Chapter 3. For large λ_n this condition means that even high quality preconditioners in $\mathcal{B}_\gamma$ can yield one-step convergence. Compare also with Lemma 3.2 where a similar condition is discussed for the components of the iteration vector.

Corollary 4.8. *If*

$$\gamma \geq \frac{\lambda_1}{\lambda_n - \lambda_1}, \tag{4.21}$$

then a preconditioner is contained in $\mathcal{B}_\gamma$ and an iteration vector can be found in $L(\lambda)$, $\lambda_1 \leq \lambda < \lambda_n$, so that PINVIT terminates in a single step within an eigenvector belonging to the smallest eigenvalue λ_1.

Proof. The convergence factor

$$\max(0, \frac{\lambda_1}{\lambda_n} - \gamma(1 - \frac{\lambda_1}{\lambda_n}))$$

given in the last row of Table 4.1 equals 0, if (4.21) is fulfilled. □

4.3 A critical comparison of convergence estimates

After having derived convergence estimates on the poorest and best convergence of PINVIT in terms of $\Phi^{\pm}_{i,j}$, see Equation (3.44), and in terms of the σ_k, cf. Table 4.1, let us now summarize the advantages and disadvantages of these representations.

$\Phi^{\pm}_{i,j}$-representation:

1. Sharp in λ_i, λ_j, λ and γ.
2. $\Phi^{+}_{i,i+1}$ in Figure 3.5 indicates superlinear convergence in each interval $[\lambda_i, \lambda_{i+1})$.
3. Main drawback: Lengthy formula $\lambda^{\pm}_{i,j}(\lambda, \gamma)$, see Equation (3.37). It is very difficult to assess its dependence on λ_i, λ_j, λ and γ. A recursive estimate, like (4.4), cannot be given.

σ_k-representation:

1. Sharp in λ_i, λ_j and γ. But only asymptotically sharp in λ; turns into a sharp estimate for $\lambda \to \lambda_i$.
2. Easy-to-use convergence factors σ_i. Dependence on λ_i, λ_j and γ is “visibly” clear.
3. Independence of λ allows a recursive representation, see for instance (4.4).

In Figure 4.1 the σ_k-convergence factors are illustrated for the same test problem as used in Section 3.4, i.e. the first eigenvalues of the 2D Laplacian on $[0, \pi]^2$. Thus this figure is to be understood as the pendant of Figure 3.5 representing the $\Phi^{\pm}_{i,j}$ factors.

Figure 4.1(a) displays the factors of extremal convergence σ_1 and σ_7 as defined by Theorem 4.6. While σ_1^2 is drawn for $\gamma = 0, 0.1, \ldots, 1$, the factor σ_7^2 is plotted only for $\gamma = 0$ and $\gamma = 1$.

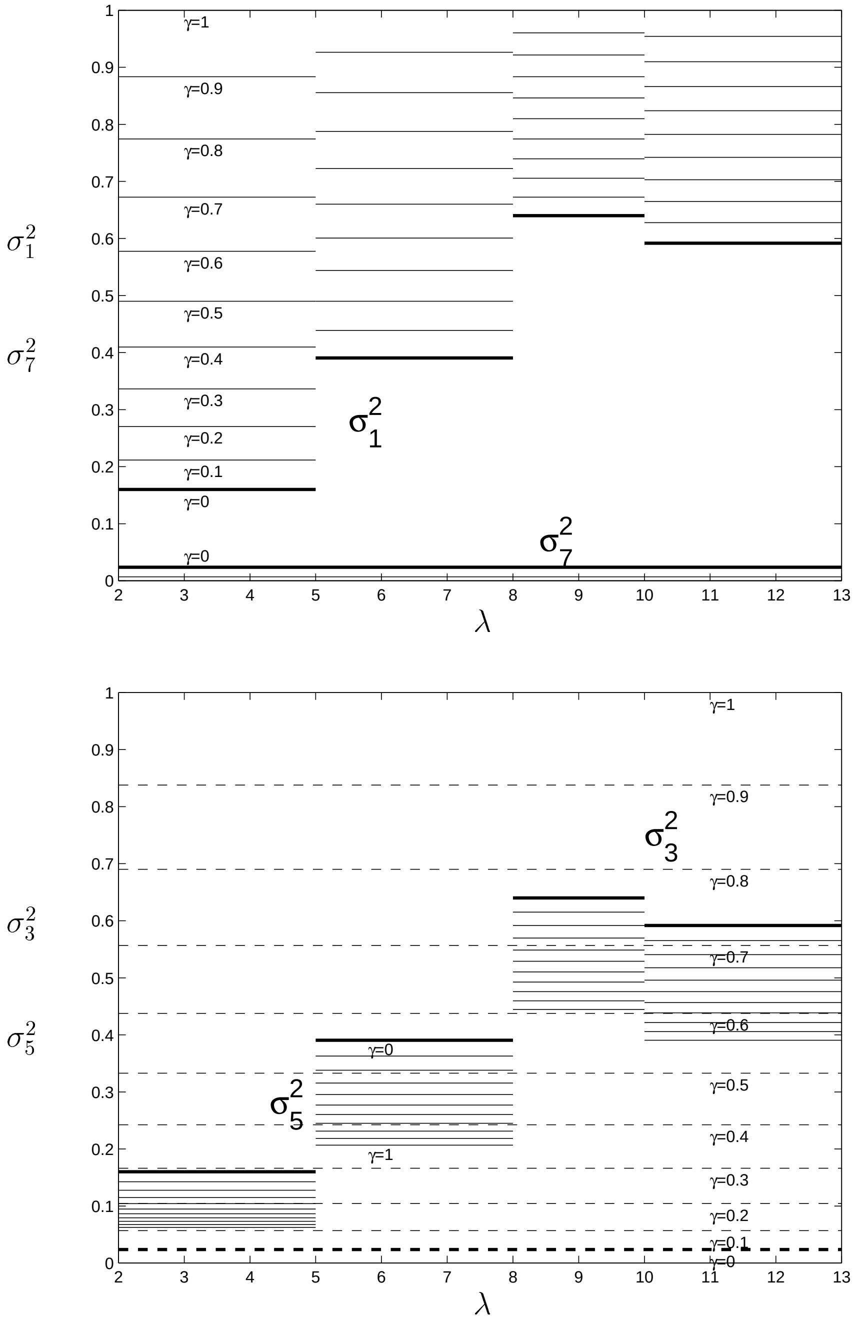


Figure 4.1: $\boxed{\begin{smallmatrix}a\\b\end{smallmatrix}}$ *PINVIT convergence factors* σ_k. *(a) Slowest (fastest) PINVIT convergence by* σ_1^2 *(*σ_7^2*) for* $\gamma = 0, 0.1, \ldots, 1.0$. *(b) Combinations of best/poorest preconditioning with poorest/best choice of* $x \in L(\lambda)$.

For $\gamma = 1$ it holds $\sigma_7^2 = 1/169$—exemplifying the excellent convergence realized by the most favorable preconditioner in the set $\mathcal{B}_1$.

The convergence factors for $\gamma = 0$, or exact preconditioning, are drawn by bold lines. These are the convergence factors of INVIT, i.e.

$$\sigma_1 = \lambda_i/\lambda_{i+1} \qquad \text{and} \qquad \sigma_7 = \lambda_1/\lambda_n,$$

as shown in Figure 1.3. Finally, in Figure 4.1(b) the factors σ_3^2 (dashed lines) and σ_5^2 (solid lines) are plotted; they correspond to the remaining combinations of best/poorest preconditioning with poorest/best choice of the iteration vector.

5. A PRECONDITIONED SUBSPACE EIGENSOLVER

Subspace iterations for computing invariant subspaces to a modest number of eigenvalues at one of the ends of the spectrum, or in the neighborhood of some shift parameter (in the case of shift-techniques) are widely accepted for non-preconditioned eigensolvers, see Parlett [107] and van der Vorst [50]. It is well known that subspace techniques are very effective in calculating clustered eigenvalues if the relative gap to the non-wanted part of the spectrum is not too small. Among others, the most prominent examples are the block (inverse) power method, see Bauer [8] and Rutishauser [116] and the (block) Lanzcos process [78]. Subspace eigensolvers for very large problems that result from mesh discretizations of partial differential operators have been described by Hackbusch [52, 55] and Mandel and McCormick [84].

A diagonally *preconditioned* subspace scheme was introduced by Davidson [27], in which the subspace dimension is increased in each step. The Davidson scheme is very popular in electronic-structure theory [57, 91] and has been extended to the very successful Jacobi-Davidson iteration method [121].

Preconditioned subspace iterations for mesh eigenproblems have been suggested and analyzed, e.g., by Samokish [119], D'yakonov and Knyazev [33, 34], Meyer [89], D'yakonov [32], Bramble, Knyazev and Pasciak [15], Zhang, Golub and Law [149], N. [98] and others. In the works of D'yakonov and Knyazev the first explicit convergence estimates independent of the meshwidths have been given for a somewhat simplified preconditioned subspace iteration scheme, see Equations (5.1) and (5.3).

5.1 Analysis of simplified subspace solvers

In order to point out some differences between the proof techniques used by D'yakonov and Knyazev [33, 34] and the one used in the more recent proof of Bramble, Knyazev and Pasciak [15] and, finally, the proof in [98], let us summarize the definition of PINVIT(1,s) from Algorithm 1.5. For a given subspace $\mathcal{S}$ of $\mathbb{R}^n$ let V be the orthogonal matrix of rank s containing in its columns the Ritz vectors v_i, $i = 1, \dots, s$, of A in $\mathcal{S}$. The diagonal matrix $\Theta \in \mathbb{R}^{s \times s}$ contains the Ritz values $\theta_1 \leq \theta_2 \leq \dots \leq \theta_s$ on its diagonal. Then PINVIT(1,s) maps V to

$$\tilde{V} = V - B^{-1}(AV - V\Theta), \tag{5.1}$$

for some preconditioner $B^{-1} \in \mathcal{B}_\gamma$ by (2.3). Subsequently, the Rayleigh-Ritz procedure

$$\tilde{V} \xrightarrow[\text{Rayleigh-Ritz}]{} (V', \Theta') \tag{5.2}$$

supplies the matrix $V' \in \mathbb{R}^{n \times s}$ containing the new Ritz vectors as well as the diagonal matrix Θ' of the new Ritz values, $\theta'_1 \leq \ldots \leq \theta'_s$.

In contrast to PINVIT(1,s) the simplified scheme analyzed by D'yakonov and Knyazev [33, 34] reads

$$\tilde{V} = V - B^{-1}(AV - \alpha V) = \big(I - B^{-1}(A - \alpha I)\big)\, V, \tag{5.3}$$

i.e. the diagonal matrix Θ in (5.1) has been substituted by the real parameter α. Therein α is identified with the largest Ritz value θ_s with respect to V. The iteration serves to determine more precise approximations to λ_s on the basis of given approximations $\theta_1, \ldots, \theta_s$ for the first s eigenvalues. Having computed a sufficiently accurate approximation to λ_s by some steps of (5.3), the iteration can be used to compute the next eigenvalue λ_{s+1} by increasing the dimension of the subspace associated with V.

The analysis of D'yakonov and Knyazev [33, 34] understands

$$R = I - B^{-1}(A - \alpha I),$$

(as factored out in (5.3)) as a fixed iteration operator acting on the subspace defined by V. Adhering to this point of view, the matrix $R_i = I - B^{-1}(A - \theta_i I)$ can be interpreted as the iteration operator for the ith Ritz vector v_i. Hence, in these early works the main difficulty was seen in the fact of having *several different* iteration operators on the approximating subspace.

Bramble, Knyazev and Pasciak [15] keep up this point of view, but overcome the difficulty of a "non-constant" iteration operator. They explicitly analyze damping properties of the iteration operator

$$R_i = I - Q_s^\perp B^{-1}(A - \theta_i I), \tag{5.4}$$

see Section 4 of [15]. Therein, R_i only acts on the Ritz vector v_i and $Q_s^\perp$ is the A-orthogonal projection onto the orthogonal complement of the s-dimensional invariant subspace to the smallest eigenvalues of A. The convergence factors in [15] contain the ratio λ_i/λ_{s+1}, as anticipated from the convergence theory of the classical subspace iteration. But as a severe drawback, Theorem 2.1 in [15] imposes restrictive conditions under which PINVIT(1,s) is guaranteed to converge to some invariant subspace of A.

In contrast to this, the interpretation of PINVIT(k,s) as approximating INVIT(k,s), see Section 1.2, does not factor out an iteration operator like (5.4), but leads to an error propagation equation for PINVIT(1,s) of the form

$$\tilde{V} - A^{-1}V\Theta = (I - B^{-1}A)(V - A^{-1}V\Theta). \tag{5.5}$$

In this equation the "iteration operator" is simply the error propagation matrix $I - B^{-1}A$ and all difficulties, as described above, disappear. Obviously, we rewrite (5.5) in the form (5.1) whenever $\tilde{V}$ is computed in practice. This view on PINVIT(1,s) has distinct advantages, making the geometric convergence analysis possible as given in [98]. In order to highlight this simple geometry, let us show how easily one can derive that PINVIT(1,s) preserves the rank of the iteration subspace; here we give a slightly generalized version of Lemma 3.1 in [98]. Lemma 5.1 should be understood as the subspace-pendant of Lemma 2.4, which also discloses the geometric interpretation of (5.6) and (5.7).

Lemma 5.1. *Let $V \in \mathbb{R}^{n\times s}$ contain in its columns s Ritz vectors of A so that the matrix of Ritz values $\Theta = V^TAV$ is diagonal. Then it holds*

$$\begin{aligned} V^TA(V - A^{-1}V\Theta) &= 0 \in \mathbb{R}^{s\times s}, && (5.6)\\ (A^{-1}V\Theta)^TA(A^{-1}V\Theta) &= V^TAV + (V - A^{-1}V\Theta)^TA(V - A^{-1}V\Theta), && (5.7)\\ \mathrm{rank}(\tilde{V}) &= \mathrm{rank}(V). && (5.8) \end{aligned}$$

Proof. Equation (5.6) follows simply by the definition of V and Θ. Thus we get the A-orthogonal decomposition

$$\begin{aligned} &V^TAV + (V - A^{-1}V\Theta)^TA(V - A^{-1}V\Theta) = \\ &\Theta - (A^{-1}V\Theta)^TA(V - A^{-1}V\Theta) = (A^{-1}V\Theta)^TA(A^{-1}V\Theta). \end{aligned}$$

Finally, to show that PINVIT(1,s) preserves the rank, let $y \in \mathbb{R}^s$ with $Vy \neq 0$. (Formally, we allow rank-deficient V.) Then

$$\begin{aligned} \|\tilde{V}y\|_A &\geq \|A^{-1}V\Theta y\|_A - \|(I - B^{-1}A)(V - A^{-1}V\Theta)y\|_A \\ &\geq \|A^{-1}V\Theta y\|_A - \|(V - A^{-1}V\Theta)y\|_A \\ &= \frac{\|A^{-1}V\Theta y\|_A^2 - \|(V - A^{-1}V\Theta)y\|_A^2}{\|A^{-1}V\Theta y\|_A + \|(V - A^{-1}V\Theta)y\|_A} \\ &= \frac{\|Vy\|_A^2}{\|A^{-1}V\Theta y\|_A + \|(V - A^{-1}V\Theta)y\|_A} > 0. \end{aligned}$$

The last inequality holds, since the numerator $\|Vy\|_A$ is nonzero by the assumption. Moreover, the denominator is positive, too, since in the case of a vanishing $A^{-1}V\Theta y$ the second summand $\|(V - A^{-1}V\Theta)y\|_A$ would remain positive. □

Note that rank preservation of PINVIT(1,s) provides the necessity to ensure that $\mathrm{rank}(V) = s$. Otherwise, the Rayleigh-Ritz procedure will produce spurious vanishing "Ritz values".

5.2 A convergence theorem

In this section we present an extended form of the central convergence theorem for PINVIT(1,s), whose basic version has been given in [98]. This improved form makes use of the simplified representation of the PINVIT convergence estimates as derived in Chapter 4.

The most surprising fact concerning the following theorem is that we encounter once again the estimates for the decrease of the Rayleigh quotient as derived for the vector scheme PINVIT. In more detail, Theorem 5.2 claims that the ith Ritz value ($i = 1, \dots, s$) in the subspace scheme decreases exactly like the Rayleigh quotient, if the *vector scheme* PINVIT is applied to the ith Ritz vector. Therefore, Theorem 5.2 does not reflect or express the accelerating influence of the Rayleigh-Ritz procedure. In contrast to this, from the classical subspace iteration applied to an s-dimensional subspace, one would anticipate a convergence factor, which is determined by a quantity like θ_i/λ_{s+1}.

Nevertheless, the described convergence behavior is not really a weakness of the convergence estimates, since the presented bounds are sharp for each Ritz value individually. The decisive point is that these estimates are not sharp *collectively*—for all Ritz values at the *same* time. We refer to Section 5.3 for the further discussion.

The preconditioned subspace eigensolver convergence theorem reads as follows:

Theorem 5.2. *Let $V = [v_1, \dots, s] \in \mathbb{R}^{n \times s}$, $s < n$, be an orthogonal matrix where the v_i are Ritz vectors of A. Application of PINVIT(1,s) defines the matrices of new Ritz vectors V' and new Ritz values Θ' by (5.1) and (5.2). Then we have:*

1. *For $\theta_i \in [\lambda_{k_i}, \lambda_{k_i+1})$ and $i = 1, \dots, s$ it holds that*
$$\theta_i' \leq \lambda_{k_i,k_i+1}(\theta_i, \gamma), \tag{5.9}$$
where $\lambda_{i,j}$ is defined by (2.27). This estimate is sharp in λ_{k_i}, λ_{k_i+1}, θ_i and γ for each i in a sense that a preconditioner B^{-1} and a subspace V can be constructed so that (5.9) is attained. But the bound (5.9) is not necessarily sharp for all Ritz values collectively.

2. *On the assumptions made above it also holds for θ_i' that either $\theta_i' < \lambda_{k_i}$ (unless $\theta_i < \lambda_{i+1}$) or that*
$$\Delta_{k_i,k_i+1}(\theta_i') \leq \left(\gamma + (1-\gamma)\frac{\lambda_{k_i}}{\lambda_{k_i+1}}\right)^2 \Delta_{k_i,k_i+1}(\theta_i), \tag{5.10}$$
where Δ_{k_i,k_i+1} is defined by (4.3). The latter estimate is sharp in λ_{k_i}, λ_{k_i+1} and γ, but only asymptotically sharp in θ_i. It turns into a sharp estimate as $\theta_i \to \lambda_{k_i}$. The estimate is not necessarily sharp for all Ritz values collectively.

The proof of estimate (5.9) is given in [98]. To prove the second estimate (5.10), just note that the bound (5.9) for a fixed index k_i is precisely the same as the one derived in Theorem

2.8 for the vector scheme PINVIT(1). Therefore, the convergence theory presented in Chapter 4 can be applied, which yields (5.10).

Let us summarize the main idea of the PINVIT(1,s) convergence proof. The obvious idea to treat PINVIT(1,s) by considering the s columns of $\tilde{V}$ independently (or in other words, to analyze PINVIT(1,s) as s separate PINVIT(1) iterations) is doomed to failure because the s balls $E_\gamma(v_i)$ are not always pairwise disjoint sets. This might make PINVIT(1,s) seem to be a rank reducing scheme, erroneously. Instead, the first step of the convergence proof consists in showing that the largest Ritz value θ_s of V behaves like the Rayleigh quotient in the PINVIT(1) scheme, i.e. if $\theta_s \in [\lambda_p, \lambda_{p+1})$, then

$$\theta_s' \leq \lambda_{p,p+1}(\theta_s, \gamma). \tag{5.11}$$

To prove the latter equation, the idea is to rewrite for nonzero $y \in \mathbb{R}^s$ the error propagation equation

$$\tilde{V}y = A^{-1}V\Theta y + (I - B^{-1}A)(V - A^{-1}V\Theta)y$$

as

$$\tilde{V}y = \lambda(z)A^{-1}z + (I - B^{-1}A)(Vy - \lambda(z)A^{-1}z) \tag{5.12}$$

for $z = \lambda(V\Theta y)^{-1}V\Theta y$. The last equation differs from PINVIT(1) applied to z in the term $z - \lambda(z)A^{-1}z$ substituted by $Vy - \lambda(z)A^{-1}z$. Both latter terms can be interpreted as the radii of the balls spanned by the set $\mathcal{B}_\gamma$. Direct computation shows that

$$\|Vy - \lambda(z)A^{-1}z\|_A \leq \|z - \lambda(z)A^{-1}z\|_A.$$

Hence the scheme (5.12) defines the smaller ball. We conclude that the Ritz value θ_s in the case of PINVIT(1,s) decreases faster than the Rayleigh quotient if PINVIT is applied to z. This finally proves (5.11). The estimates for the remaining Ritz values of the approximating subspace are derived by induction on the subspace dimension together with the Courant-Fischer principles.

5.3 Generalizations and remarks

Because of its practical importance, particularly for finite element discretizations of self-adjoint and coercive elliptic differential operators, let us mention the generalized matrix eigenvalue problem. Theorem 5.2 also holds for the generalized eigenvalue problem

$$Ax = \lambda Mx$$

with symmetric positive definite matrices A and M, see [73, 97]. Then the error propagation equation reads

$$\tilde{V} - A^{-1}MV\Theta = (I - B^{-1}A)(V - A^{-1}MV\Theta), \tag{5.13}$$

where the error propagation matrix is factored out on the right-hand side, which explains that for the preconditioner the assumption (2.2) remains valid. Consequently, the iterative scheme of PINVIT(1,s) is given by

$$\tilde{V} = V - B^{-1}(AV - MV\Theta). \tag{5.14}$$

If we relinquish positive definiteness of M, convergence estimates for a slightly modified scheme (a scaling parameter is introduced and M is substituted by a shifted matrix) are available, see [73]. The proof technique was originally suggested by Knyazev [68], see also [32].

To summarize, we reproduce from [73] the advantages of the subspace convergence theory given in [98].

- Convergence to an invariant subspace is guaranteed for any initial subspace.
- The convergence rate estimate can be applied recursively.
- The convergence estimate for each of the s Ritz values of PINVIT(1,s) is exactly the same as the one given in Theorem 2.8 for PINVIT(1).
- The estimates are individually sharp in a sense that for each Ritz value an initial subspace and a preconditioner can be constructed so that the estimate is attained.

The only serious drawback of these estimates is that they do not reflect the typical behavior as known from subspace iteration [107], namely that the convergence of the ith Ritz value is controlled by the ratio λ_i/λ_{s+1}, where λ_{s+1} is the first unwanted eigenvalue. This lack of the PINVIT(1,s) estimates takes effect particularly if the eigenvalues of interest $\lambda_1, \ldots, \lambda_s$ include a cluster of eigenvalues so that by nature of estimate (5.9) the convergence of the eigenvalues within the cluster deteriorates as the degree of clustering increases.

The reason for this behavior is to be seen in the fact that the Theorem 5.2 makes no *collective* assumption on the quality of the approximating *subspace* and that by no means $\theta_s \leq \lambda_{s+1}$ is guaranteed. Instead, Inequality (5.9) only requires the indexes k_i and k_{i+1} of the eigenvalues enclosing each of the Ritz values. Inasmuch as the condition $\theta_s \leq \lambda_{s+1}$ is not fulfilled, closeness of θ_i to some eigenvalue λ_j does not imply that the Ritz vector v_i approximates the eigenvector belonging to λ_j. Hence, some Ritz vectors approximating specific eigenvectors very well, may be mixed with those of poor quality. In such a subspace one cannot expect to have the typical λ_i/λ_{s+1} convergence factor. A remedy against this situation would be an assumption on the quality of the initial subspace like Inequality (2.4) in [15] or some other comparable condition on the angle enclosed by the actual subspace and the wanted invariant subspace.

6. PINVIT(2) – Preconditioned steepest descent

So far we have analyzed the basic preconditioned eigensolvers PINVIT(1) and the corresponding subspace scheme PINVIT(1,s). For both schemes sharp non-asymptotic convergence estimates have been presented. These estimates are upper bounds for the decrease of the Rayleigh quotients of the iterates or for the decrease of the Ritz values belonging to the actual iteration subspace, respectively. In other words, we have now completely described the level $k = 1$ within the hierarchy of preconditioned eigensolvers PINVIT(k,s), as introduced in Algorithm 1.5.

Let us now proceed with the analysis of the more complex scheme PINVIT(2), or by using the customary naming, Preconditioned Steepest Descent. This eigensolver involves the application of the Rayleigh-Ritz method to the 2D subspace spanned by the actual iterate and its preconditioned residual. As a first trivial result, the Courant-Fischer principles ensure that this scheme converges at least as fast as PINVIT(1), cf. Lemma 1.7.

This chapter is organized as follows: In Section 6.1 we introduce PINVIT(2) and show by elementary examples that it may converge much more rapidly than PINVIT(1). Moreover, we define a line of demarcation to steepest descent methods for the eigenvalue problem. We highlight that *steepest descent* and *preconditioned steepest descent*, in spite of their common roots of naming, are only weakly related. As a result of this discussion, we see a much closer relation between PINVIT(2) and INVIT(2).

Therefore, and as a first step toward a convergence analysis of PINVIT(2), a new convergence analysis of INVIT(2) providing sharp convergence estimates is given in Section 6.2. In Section 6.3 a convergence analysis of PINVIT(2) follows, which is, once more, mainly founded upon the underlying geometry. While the dependence of poorest convergence of PINVIT(2) on the choice of the preconditioner is cleared up completely, a corresponding mini-dimensional analysis of PINVIT(2) is based upon a conjecture of 3D-extremal convergence. Finally, in Section 6.5 some numerical illustration is given, which also provides numerical evidence for the validity of the 3D-conjecture.

6.1 Rayleigh-Ritz accelerates convergence

An obvious and simple way to improve convergence of PINVIT(1) consists in scaling the preconditioned residual by some parameter ω, i.e.

$$x'(\omega) = x - \omega B^{-1}(Ax - \lambda x), \tag{6.1}$$

and to determine ω in such a way that the Rayleigh quotient of $x'(\omega)$ is minimized. Such a minimization appears as the natural choice of ω, since we measure the convergence of PINVIT(1) by the decrease of the Rayleigh quotient achievable per step. The Rayleigh quotient of the new iterate reads

$$\lambda(x') = \lambda(x - \omega^* d) = \min_{\omega\in\mathbb{R}} \frac{(x - \omega d, A(x - \omega d))}{(x - \omega d, x - \omega d)}, \tag{6.2}$$

where $d = B^{-1}(Ax - \lambda x)$ denotes the preconditioned residual. The optimal scaling constant ω^* can easily be determined by differentiating the Rayleigh quotient of (6.1) by ω; the necessary condition for a relative extremum yields a second order polynomial in ω^*. The resulting scheme (6.1) in ω^* is called Preconditioned Steepest Descent [71] or more systematically PINVIT(2) within the classification given in Section 1.2.

Non-asymptotic convergence estimates for PINVIT(2) are unknown so far in spite of long-standing efforts, cf. [71, 73]. The idea of preconditioned steepest descent/ascent methods is discussed in Kantorovich [65, 66]. An asymptotically sharp estimate (sharp for $\lambda(x')$ tending to λ_1) was given by Samokish; cf. Equation (10) in [119]. See also Godunov, Ogneva and Prokopov [46], Knyazev [69], estimates (3.9)–(3.12), as well as the monograph of D’yakonov [32, Sections 9.4.1 and 9.4.4] containing additional references to the literature. Recently, the method of successive eigenvalue relaxation has been presented by Ovtchinnikov and Xanthis [105, 106], a scheme which relies on consecutive relaxation steps in the directions of the preconditioned residuals belonging to the Ritz vectors of the actual approximating subspace.

Here, we prefer an alternative but equivalent representation of (6.1) and (6.2) based on the Rayleigh-Ritz procedure applied to the 2D column space of $V = [x, d]$. In order to find the minimal Rayleigh quotient in $\mathrm{span}\{x, d\}$ one has to solve the 2×2 generalized eigenvalue problem

$$\bar{A}U = \bar{M}U\Theta, \qquad \Theta = \mathrm{diag}(\theta_1, \theta_2),$$

with

$$\bar{A} = V^T A V = \begin{pmatrix} (x, Ax) & (d, Ax) \\ (d, Ax) & (d, Ad) \end{pmatrix} \tag{6.3}$$

and

$$\bar{M} = V^T V = \begin{pmatrix} (x, x) & (d, x) \\ (d, x) & (d, d) \end{pmatrix}. \tag{6.4}$$

The smaller Ritz value θ_1 equals $\lambda(x')$ and reads

$$\theta_1 = \frac{\tau}{2\det(M)} - \sqrt{\frac{\tau^2}{4(\det(M))^2} - \frac{\det(A)}{\det(M)}},$$

for

$$\tau = (x, Ax)(d, d) + (x, x)(d, Ad) - 2(d, Ax)(d, x).$$

The discriminant is positive since the Ritz values are real numbers larger than λ_1. Denote by v the eigenvector of $(\bar{A}, \bar{M})$ belonging to θ_1. Then x' is collinear to $Vv = xv_1 + dv_2$. Obviously, ω^* equals the ratio of the components of v,

$$\omega^* = \frac{v_2}{v_1}, \tag{6.5}$$

as long as $v_1 \neq 0$. But the denominator of (6.5) may vanish if the preconditioned residual d is a collinear vector to the eigenvector x_1. We get rid of this singularity later in Lemma 6.15 by scaling the iterate x instead of the preconditioned residual. Note that the numerator of (6.5) is nonzero as long as x is different from an eigenvector of A; otherwise our preconditioned eigensolvers would be stationary and their application would make no sense.

Before we start analyzing the convergence of PINVIT(2), let us treat the question of whether or not such an attempt is worthwhile. Can we expect improved convergence estimates for PINVIT(2) concerning the poorest convergence compared to that of PINVIT(1)? A simple argument reveals a positive answer, at least concerning the bounds for the Rayleigh quotient of that type presented in Theorem 2.8. First one should observe that $\theta_1 \leq \lambda(x - 1d)$ so that PINVIT(2) does not converge more slowly than PINVIT(1). The second and decisive argument is that the poorest convergence of PINVIT(1), in the domain $[\lambda_i, \lambda_{i+1}]$ of Rayleigh quotients, is known to be taken in the 2D space $\mathcal{S}_{i,i+1} = \operatorname{span}\{x_i, x_{i+1}\}$ as shown in [96]. Therein, x_i and x_{i+1} are the eigenvectors belonging to λ_i and λ_{i+1}. But PINVIT(2) behaves very differently in the same space $\mathcal{S}_{i,i+1}$. If the actual iterate and its preconditioned residual are both contained in $\mathcal{S}_{i,i+1}$, PINVIT(2), by the Courant-Fischer principles, converges immediately to the eigenpair (x_i, λ_i), which means a considerable acceleration of convergence.

Keeping this simple result in mind, we will not be surprised by Theorem 6.3 and Conjecture 6.16 claiming that poorest INVIT(2) as well as PINVIT(2) convergence is taken in $\operatorname{span}\{x_i, x_{i+1}, x_n\}$, i.e. the 3D space spanned by $\mathcal{S}_{i,i+1}$ and x_n. The dependence on the eigenpair (x_n, λ_n) may at a first glance appear disadvantageous, as λ_n is usually extremely large for mesh discretizations. But this dependence must be a very weak one, since even the slower convergent PINVIT(1) scheme is known to be equipped with grid-independent convergence estimates.

Finally, let us mention that the $\Delta_{p,q}$ factors for (P)INVIT(2) reflect some acceleration in comparison to those of (P)INVIT(1), too, as we will see in the following.

6.1.1 Steepest descent for the eigenproblem

In preparation of a convergence analysis for INVIT(2) let us throw a glance at the *steepest descent* for the eigenproblem. For given nonzero $x \in \mathbb{R}^n$ having the Rayleigh quotient λ, the idea of steepest descent is to correct x in the direction of the negative gradient of the Rayleigh quotient, i.e.

$$x \quad \longrightarrow \quad x' := x - \omega(Ax - \lambda x), \tag{6.6}$$

with ω minimizing the Rayleigh quotient of x'. We make use of the fact that the residual $Ax - \lambda x$ and the gradient of $\lambda(\cdot)$ are collinear vectors. Obviously, for a nonvanishing residual (6.6) succeeds in decreasing the Rayleigh quotient and the sequence of Rayleigh quotients (slowly) converges to some eigenvalue, while the vector-iterates tend to an eigenvector.

Steepest ascent for the eigenproblem, a technique for the computation of the largest eigenvalue together with an eigenvector, derives from (6.6) by choosing ω in such a way that the Rayleigh quotient of x' is *maximized*. The classical asymptotic estimates of the convergence rate of steepest descent/ascent go back to Kantorovich [65, 66] as well as to Hestenes and Karush [58]. Non-asymptotic estimates are given by Prikazchikov [111], Zhuk and Bondarenko [150] as well as by Knyazev and Skorokhodov [69, 76].

As has already been pointed out in Section 1.2, *steepest descent* and *preconditioned steepest descent*, here called PINVIT(2), are only weakly related. Only if *no* preconditioning is applied, i.e. $B = I$, PINVIT(2) will reduce to steepest descent for the Rayleigh quotient. We lay special emphasis on noting that the choice $B = I$ is way out from the usual assumption (2.2) on the quality of the preconditioner. Just note that for $B = I$ the spectral radius of the error propagation matrix $||I - A||_A$ behaves asymptotically like the largest eigenvalue of A.

But there is a remarkable, much stronger relation brought about for the choice $B^{-1} = A^{-1}$ of exact preconditioning or $\gamma = 0$, cf. Section 1.2. While PINVIT(1) for $B = A$ equals INVIT(1), the scheme PINVIT(2) for $B = A$, in contrast to (6.6), results in INVIT(2), i.e.

$$(x, \lambda) \quad \longrightarrow \quad (x' = x - \omega(x - \lambda A^{-1}x), \lambda' = \lambda(x')),$$

where ω minimizes the Rayleigh quotient of x'. Naturally, we prefer to rewrite this as

$$(x, \lambda) \quad \longrightarrow \quad (x' = x - \omega A^{-1}x, \lambda' = \lambda(x')), \tag{6.7}$$

with, once more, ω minimizing the Rayleigh quotient of x'.

In other words, preconditioned steepest descent for a decreasing γ improves to approximate INVIT(2), providing the justification to refer to (6.1) and (6.2) as PINVIT(2).

Nevertheless, the convergence analysis of the steepest *ascent* method as given by Knyazev and Skorokhodov [76] can be successfully extended to a convergence theory of INVIT(2). This is done in Section 6.2, yielding asymptotically sharp convergence estimates in terms of the $\Delta_{p,q}$ factors and for the acute angle enclosed with the eigenvector to the smallest eigenvalue. Before doing this, the result of Knyazev and Skorokhodov [76] is reformulated in Corollary 6.1

for the steepest descent method (6.6). The proof of this reformulation is trivial; just substitute A by $-A$ in Theorem 2.2 in [76].

Corollary 6.1 (Convergence of steepest descent). *If $\lambda(x) < \lambda_2$ then for the x' by (6.6) we have*

$$\frac{\Delta_{1,2}(\lambda')}{\Delta_{1,2}(\lambda)} \leq \left(\frac{1-\xi}{1+\xi}\right)^2, \tag{6.8}$$

where

$$\Delta_{1,2}(\kappa) = \frac{\kappa - \lambda_1}{\lambda_2 - \kappa} \tag{6.9}$$

and

$$\xi = (\lambda_2 - \lambda_1)/(\lambda_n - \lambda_1). \tag{6.10}$$

Moreover, for any initial vector $x^{(0)}$ with $\tan\varphi_0 < \infty$ one has

$$\frac{\tan^2\varphi_k}{\tan^2\varphi_0} \leq (1-\xi)^{2k}, \tag{6.11}$$

where φ_k denotes the acute angle between $x^{(k)}$ and x_1. The estimates are asymptotically sharp for some sequence of vectors with a Rayleigh quotient tending to λ_1.

Remark 6.2. *The estimate (6.8) together with (6.10) proves steepest descent unsuitable for matrices with large λ_n, e.g. for mesh eigenproblems, since $\lim_{\lambda_n\to\infty}\xi = 0$ implies that the convergence factor on the right-hand side of (6.8) tends to 1.*

Contrastingly, the convergence factor of INVIT(2) is bounded away from 1 for $\lambda_n \to \infty$, see Theorem 6.3.

6.2 Convergence analysis of INVIT(2)

As the first step toward a convergence analysis of PINVIT(2) we analyze the case of exact preconditioning, i.e. convergence estimates are derived for INVIT(2) as given by (6.7).

Theorem 6.3. *Let $x^{(0)} \in \mathbb{R}^n$ and*

$$\xi = \frac{\lambda_2 - \lambda_1}{\lambda_2 - \frac{\lambda_1\lambda_2}{\lambda_n}}. \tag{6.12}$$

If $\lambda_1 \leq \lambda(x^{(0)}) < \lambda_2$, then for the kth iterate $x^{(k)}$ of INVIT(2) (6.7) it holds that

$$\frac{\Delta_{1,2}(\lambda(x^{(k)}))}{\Delta_{1,2}(\lambda(x^{(0)}))} \leq \left(\frac{1-\xi}{1+\xi}\right)^{2k}, \quad \textit{for} \quad k = 1, 2, \ldots \tag{6.13}$$

This estimate is sharp in λ_1, λ_2 and λ_n. It is asymptotically sharp in λ and turns into a sharp estimate as $\lambda \to \lambda_1$.

Moreover, for any $x^{(0)}$ with $\tan\varphi_0 < \infty$ one has

$$\frac{\tan^2\varphi_k}{\tan^2\varphi_0} \leq (1-\xi)^{2k}, \tag{6.14}$$

where φ_k is the acute angle enclosed by $x^{(k)}$ and x_1.

Remark 6.4. *Formally, we obtain Equation (6.12) from (6.10) by replacing λ_i by $1/\lambda_i$ for $i = 1, 2, n$. Therefore, one might believe that INVIT(2) could be understood as steepest descent for A^{-1}. But this is in fact not true, since in both cases the Rayleigh-Ritz approximations are computed with respect to A. In contrast to that, a complete substitution of A by A^{-1} would lead to the so-called harmonic Ritz values/vectors, but Theorem 6.3 is formulated in terms of the standard Ritz approximations. Nevertheless, the proof of Theorem 6.3 tightly follows the ideas of Knyazev and Skorokhodov [76], but at various points alterations take effect which result from having A^{-1} as the "iteration operator" and from computing the Rayleigh-Ritz approximations with respect to A.*

Remark 6.5. *Since ξ by Equation (6.12) is a decreasing function in λ_n ($\lambda_n > \lambda_2$) and $(1-\xi)/(1+\xi)$ decreases in $\xi \in [1-\lambda_1/\lambda_2, 1)$, we obtain a simplified convergence factor $\tilde{\sigma}$ by*

$$\tilde{\sigma} := \lim_{\lambda_n\to\infty} \frac{1-\xi}{1+\xi} = \frac{\lambda_1}{\lambda_2 + (\lambda_2-\lambda_1)} > \frac{1-\xi}{1+\xi}, \tag{6.15}$$

which holds globally for INVIT(2). Compare with Remark 6.2 to see that $\tilde{\sigma}$ in contrast to the convergence factor of INVIT(2) is bounded away from 1 for $\lambda_n \to \infty$. We do not lose much quality of this bound by taking the limit $\lambda_n \to \infty$ whenever $\lambda_n \gg \lambda_2$ (which is usually the case for mesh eigenproblems of elliptic partial differential operators). For instance taking the discrete Laplacian Δ_h we have $\lambda_n \sim h^{-2}$ so that $\xi = 1 - \lambda_1/\lambda_2 + \mathcal{O}(h^2)$.

The comparison with the convergence factor λ_1/λ_2 of the standard INVIT(1), see Corollary 1.10, points out the impact of INVIT(2), which consists in the additional summand $\lambda_2 - \lambda_1$ hastening convergence.

Remark 6.6. *It is easy to generalize Theorem 6.3 to the case that $\lambda(x) \in [\lambda_i, \lambda_{i+1})$ for $i = 1, \ldots, n-2$. Then for the new INVIT(2) iterate x' it holds either $\lambda(x') < \lambda_i$ or*

$$\frac{\Delta_{i,i+1}(\lambda(x'))}{\Delta_{i,i+1}(\lambda(x))} \leq \left(\frac{1-\xi}{1+\xi}\right)^2 \tag{6.16}$$

with

$$\xi = \frac{\lambda_{i+1}-\lambda_i}{\lambda_{i+1} - \frac{\lambda_i\lambda_{i+1}}{\lambda_n}}. \tag{6.17}$$

This estimate is sharp in λ_i, λ_{i+1} and λ_n. It is asymptotically sharp in λ and turns into a sharp estimate as $\lambda \to \lambda_i$. In the remaining interval $[\lambda_{n-1}, \lambda_n)$ the estimate $\lambda(x') \leq \lambda_{n-1}$ is sharp by Lemma 6.12.

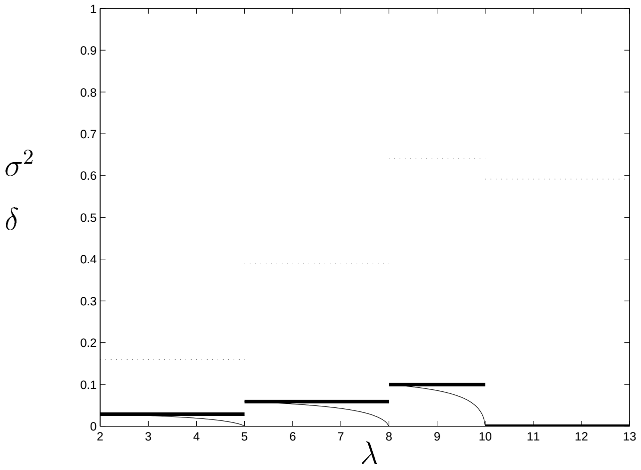


Figure 6.1: Comparison of INVIT(1) and INVIT(2) convergence estimates. Dotted lines: $\sigma^2[INVIT(1)]$. Bold solid lines: $\sigma^2[INVIT(2)]$. Solid curves $\delta := \Delta_{i,i+1}(\lambda')/\Delta_{i,i+1}(\lambda)$.

These estimates are illustrated by using the same example as in Figure 1.2, see the model analysis of INVIT(1) in Section 1.4. For the sake of comparison, Figure 6.1 shows the convergence factors $\sigma^2_{i,i+1}$ of INVIT(1) as dotted lines while the corresponding factors of INVIT(2) are drawn as bold solid lines which reflect the fact that INVIT(2) converges much more rapidly. Finally, the left-hand side of 6.16, i.e.

$$\delta := \Delta_{i,i+1}(\lambda')/\Delta_{i,i+1}(\lambda)$$

are plotted in the intervals $[\lambda_i, \lambda_{i+1})$, $i = 1, \ldots, n-2$, as solid curves illustrating the asymptotic sharpness of (6.16) as $\lambda \to \lambda_i$. See Sections 6.5 and 6.5.3 on how to compute δ.

Let us start with the convergence analysis of INVIT(2): Since ξ is independent of k it suffices to give the proof only for $n = 1$. We start with some introductory definitions and define the two subspaces

$$K^{[-1]} = \operatorname{span}\{x, A^{-1}x\}, \qquad H^{[-3]} = \operatorname{span}\{x_1, x, A^{-1}x\},$$

where x_1 denotes the eigenvector to the smallest eigenvalue λ_1. The proof is based on a mini-dimensional analysis (a concept introduced in [76]) where the convergence estimates derived in $H^{[-3]}$ turn out to hold in the general case. We exclude the trivial case $\dim K^{[-1]} = 1$ (just observe that this would mean stationarity of inverse iteration and x would be an eigenvector) as well as $\dim H^{[-3]} = 2$ since then $x_1 \in K^{[-1]}$ and $x' = x_1$. The estimates (6.13) and (6.14) would hold trivially in these cases. Next we introduce the orthogonal projector $P^{[-3]}$ on $H^{[-3]}$ and define

$$A^{[3]} := P^{[-3]}A, \qquad A^{[-3]} := P^{[-3]}A^{-1}.$$

Then $A^{[3]}$ and $A^{[-3]}$ are symmetric operators on $H^{[-3]}$ having $H^{[-3]}$ as an invariant subspace.

Lemma 6.7. *Let*

$$x' = x - \omega A^{-1}x, \quad \omega \textit{ minimizing } \frac{(x', Ax')}{(x', x')},$$

$$\tilde{x} = x - \omega A^{[-3]}x, \quad \omega \textit{ minimizing } \frac{(\tilde{x}, A^{[3]}\tilde{x})}{(\tilde{x}, \tilde{x})}.$$

Then $x' = \tilde{x}$ so that we can analyze INVIT(2) in the 3D subspace $H^{[-3]}$.

Proof. First observe that both methods span the same subspace, i.e. $K^{[-1]} = \mathrm{span}\{x, A^{[-3]}x\}$. Furthermore, the Rayleigh quotients $\lambda_A(\cdot)$ and $\lambda_{A^{[3]}}(\cdot)$ coincide on $H^{[-3]}$ since for $y \in H^{[-3]}$ it holds

$$(y, Ay) = (y, P^{[-3]}Ay) = (y, A^{[3]}y).$$

We conclude that both methods will find the same Ritz vector contained in $H^{[-3]}$. □

Let $\theta_1 \leq \theta_2 \leq \theta_3$ be the Ritz values of A with respect to the space $H^{[-3]}$ and denote by v_i the corresponding orthonormal Ritz vectors. Since $x_1 \in H^{[-3]}$ we have $\theta_1 = \lambda_1$ and $v_1 = x_1$. Then the v_i are the eigenvectors and the θ_i are the eigenvalues of $A^{[3]}$. Moreover, by the assumption $\lambda_1 < \lambda_2$ together with the Courant-Fischer theorem we have

$$\theta_1 = \lambda_1 < \lambda_2 \leq \theta_2 \leq \theta_3 \leq \lambda_n. \tag{6.18}$$

To show that $\theta_2 = \theta_3$ embodies a trivial case, represent x (which is at present assumed not to be an eigenvector of A) in the form

$$x = \alpha x_1 + \beta v, \qquad \text{for some } v \in \mathrm{span}\{v_2, v_3\}.$$

Because of $\dim H^{[-3]} = 3$ we have $\beta \neq 0$. If $\theta_2 = \theta_3$, then v would be an eigenvector of $A^{[3]}$ so that $\alpha \neq 0$. Hence $K^{[-1]} = \mathrm{span}\{x_1, v\}$ and (6.13) holds trivially since $K^{[-1]} = H^{[-3]}$ and $\dim H^{[-3]} = 2$. In the following we suppose $\theta_2 < \theta_3$.

Next we give the justification that it suffices to prove only the counterpart of the convergence estimate (6.13) in $H^{[-3]}$. Therefore let

$$\kappa = \frac{\theta_1(\theta_3 - \theta_2)}{\theta_2(\theta_3 - \theta_1)} \tag{6.19}$$

and

$$\Delta_{1,2}^{[-3]}(\lambda') = \frac{\lambda' - \theta_1}{\theta_2 - \lambda'}, \qquad \Delta_{1,2}^{[-3]}(\lambda) = \frac{\lambda - \theta_1}{\theta_2 - \lambda}. \tag{6.20}$$

We will show in the following

$$\frac{\Delta_{1,2}(\lambda')}{\Delta_{1,2}(\lambda)} \leq \frac{\Delta_{1,2}^{[-3]}(\lambda')}{\Delta_{1,2}^{[-3]}(\lambda)} \leq \frac{\kappa^2}{(2-\kappa)^2} \leq \left(\frac{1-\xi}{1+\xi}\right)^2. \tag{6.21}$$

The far left inequality follows from $\lambda_1 = \theta_1$ and

$$\frac{\theta_2 - \lambda}{\theta_2 - \lambda'} \geq \frac{\lambda_2 - \lambda}{\lambda_2 - \lambda'},$$

whereas the far right inequality is a consequence of $\kappa \leq 1 - \xi$. This last inequality follows from (6.18) resulting in

$$\kappa = \frac{\lambda_1(\theta_3 - \theta_2)}{\theta_2(\theta_3 - \lambda_1)} \leq \frac{\lambda_1(\theta_3 - \lambda_2)}{\lambda_2(\theta_3 - \lambda_1)} \leq \frac{\lambda_1(\lambda_n - \lambda_2)}{\lambda_2(\lambda_n - \lambda_1)} = 1 - \xi.$$

We infer that the “mini-dimensional”’ estimate

$$\frac{\Delta_{1,2}^{[-3]}(\lambda')}{\Delta_{1,2}^{[-3]}(\lambda)} \leq \frac{\kappa^2}{(2-\kappa)^2}$$

is not weaker than the corresponding estimate in the $\mathbb{R}^n$. The remaining part of the analysis is restricted to $H^{[-3]}$.

We denote by $\tilde{\theta}_1$, $\tilde{\theta}_2$ the Ritz values of $K^{[-1]}$ in $H^{[-3]}$. (By Lemma 6.7 they agree with the Ritz values in $\mathbb{R}^n$.) Since $K^{[-1]} \subset H^{[-3]}$ it holds

$$\theta_1 \leq \tilde{\theta}_1 \leq \theta_2 \leq \tilde{\theta}_2 \leq \theta_3.$$

Next we show that the only nontrivial case is

$$\theta_1 < \tilde{\theta}_1 < \theta_2 < \tilde{\theta}_2 < \theta_3.$$

Lemma 6.8. *Let x be expanded in the Ritz vectors v_i*

$$x = \sum_{i=1}^{3} a_i v_i.$$

Even if one $\tilde{\theta}_i$ equals some θ_j ($j = 1, 2, 3$), then $a_1 a_2 a_3 = 0$ and the convergence estimates (6.13), (6.14) hold trivially.

Proof. For the Ritz values $\tilde{\theta}_1$ and $\tilde{\theta}_2$ it holds

$$\det \begin{pmatrix} m_1 - \tilde{\theta} m_0 & m_0 - \tilde{\theta} m_{-1} \\ m_0 - \tilde{\theta} m_{-1} & m_{-1} - \tilde{\theta} m_{-2} \end{pmatrix} = \tilde{\theta}^2 q_2 + \tilde{\theta} q_1 + q_0 = 0$$

with $m_l = (x, (A^{[3]})^l x) = \sum_{i=1}^{3} \theta_i^l a_i^2$. We obtain the coefficients q_i in their symmetrized form (write down q_i by its definition and add the same term with interchanged indexes of summation)

$$\begin{aligned} q_2 &= \frac{1}{2}\sum_{i,j=1}^{3}(\theta_i^{-1}-\theta_j^{-1})^2 a_i^2 a_j^2, \\ q_1 &= \frac{1}{2}\sum_{i,j=1}^{3}-(\theta_i+\theta_j)(\theta_i^{-1}-\theta_j^{-1})^2 a_i^2 a_j^2, \\ q_0 &= \frac{1}{2}\sum_{i,j=1}^{3}\theta_i\theta_j(\theta_i^{-1}-\theta_j^{-1})^2 a_i^2 a_j^2. \end{aligned}$$

This results in

$$\tilde{\theta}_1+\tilde{\theta}_2 = \frac{\sum_{i,j=1}^{3}(\theta_i+\theta_j)(\theta_i^{-1}-\theta_j^{-1})^2 a_i^2 a_j^2}{\sum_{i,j=1}^{3}(\theta_i^{-1}-\theta_j^{-1})^2 a_i^2 a_j^2} \tag{6.22}$$

$$\tilde{\theta}_1\tilde{\theta}_2 = \frac{\sum_{i,j=1}^{3}\theta_i\theta_j(\theta_i^{-1}-\theta_j^{-1})^2 a_i^2 a_j^2}{\sum_{i,j=1}^{3}(\theta_i^{-1}-\theta_j^{-1})^2 a_i^2 a_j^2} \tag{6.23}$$

Let us first assume $\tilde{\theta}_2=\theta_3$. Eliminating $\tilde{\theta}_1$ from (6.22) and (6.23) results in

$$\frac{q_2\theta_3^2+q_1\theta_3+q_0}{q_2\theta_3}=0,$$

which implies

$$\sum_{i,j=1}^{3}\left\{\theta_3(\theta_i+\theta_j)-\theta_3^2-\theta_i\theta_j\right\}(\theta_i^{-1}-\theta_j^{-1})^2 a_i^2 a_j^2=0.$$

Only a_1a_2 has a nonvanishing coefficient so that $a_1a_2=0$. In a similar way we show that $\tilde{\theta}_2=\theta_2$ implies $a_1a_3=0$. Finally, $\tilde{\theta}_1=\theta_1$ ($\tilde{\theta}_1=\theta_2$) implies $a_2a_3=0$ ($a_1a_3=0$).

Assuming $a_1a_2a_3=0$ always means $a_2a_3=0$ since $a_1=0$ is excluded by the assumption $\lambda(x)<\lambda_2$. If $a_2a_3=0$ then $x\in\operatorname{span}\{x_1,v_2\}$ or $x\in\operatorname{span}\{x_1,v_3\}$. In either case $x'=x_1$ or $\lambda'=\lambda_1$. □

Proof of Theorem 6.3. Now we consider the nontrivial case

$$\theta_1<\tilde{\theta}_1<\theta_2<\tilde{\theta}_2<\theta_3$$

and represent x in the form

$$x=x_1+\alpha_0v_2+\beta_0v_3, \tag{6.24}$$

where, for convenience, x is normalized in a way that the coefficient of x_1 equals 1. By expanding the Rayleigh quotient as a function of α_0 and β_0 we obtain

$$\Delta_{1,2}^{[-3]}(\lambda)=\frac{\alpha_0^2(\theta_2-\theta_1)+\beta_0^2(\theta_3-\theta_1)}{\theta_2-\theta_1-\beta_0^2(\theta_3-\theta_2)} \tag{6.25}$$

and $\Delta_{1,2}^{[-3]}(\lambda') = (\tilde{\theta}_1 - \theta_1)/(\theta_2 - \tilde{\theta}_1)$. In order to express α_0 and β_0 in terms of θ_i, $\tilde{\theta}_i$ let $c \in H^{[-3]}$ be a vector orthogonal to $K^{[-1]}$ with $c = \sum_{i=1}^{3} c_i v_i$, $\|c\| = 1$. First we represent c as a function of x

$$c = \frac{(A - \tilde{\theta}_1 I)(A - \tilde{\theta}_2 I)A^{-1}x}{\|(A - \tilde{\theta}_1 I)(A - \tilde{\theta}_2 I)A^{-1}x\|}.$$

The last equation is established by setting $A^{-1}x = y$ and recognizing that $(A - \tilde{\theta}_1 I)(A - \tilde{\theta}_2 I)y$ is a vector perpendicular to the Krylov space $\mathcal{K}^2 = \mathrm{span}\{y, Ay\}$, see Lemma 12.3.1 in [107]. Additionally, we have a representation of c depending on x'

$$c = \frac{(A - \tilde{\theta}_1 I)x'}{\|(A - \tilde{\theta}_1 I)x'\|},$$

since the residual of the Ritz vector x' is orthogonal to $K^{[-1]}$. As a third condition it holds for the components of c

$$c_i^2 = \frac{\prod_{j=1}^{2}(\tilde{\theta}_j - \theta_j)}{\prod_{j=1, j\neq i}^{3}(\theta_j - \theta_j)},$$

see Section 12.6.2 in [48]. A rather wearisome calculation leads to the longish formula

$$\alpha_0^2 = \frac{(\tilde{\theta}_1 - \theta_1)(\tilde{\theta}_2 - \theta_1)(\theta_3 - \theta_1)\theta_2^2}{(\theta_2 - \tilde{\theta}_1)(\tilde{\theta}_2 - \theta_2)(\theta_3 - \theta_2)\theta_1^2}, \qquad \beta_0^2 = \frac{(\tilde{\theta}_1 - \theta_1)(\tilde{\theta}_2 - \theta_1)(\theta_2 - \theta_1)\theta_3^2}{(\theta_3 - \tilde{\theta}_1)(\theta_3 - \tilde{\theta}_2)(\theta_3 - \theta_2)\theta_1^2},$$
$$\alpha_1^2 = \frac{(\tilde{\theta}_1 - \theta_1)(\tilde{\theta}_2 - \theta_2)(\theta_3 - \theta_1)}{(\theta_2 - \tilde{\theta}_1)(\tilde{\theta}_2 - \theta_1)(\theta_3 - \theta_2)}, \qquad \beta_1^2 = \frac{(\tilde{\theta}_1 - \theta_1)(\theta_3 - \tilde{\theta}_2)(\theta_2 - \theta_1)}{(\theta_3 - \tilde{\theta}_1)(\tilde{\theta}_2 - \theta_1)(\theta_3 - \theta_2)}.$$

The following notation turns out as useful

$$\mu_1 = \frac{\theta_1(\theta_3 - \tilde{\theta}_1)}{\theta_2(\theta_3 - \theta_1)} \in (\kappa, \frac{\theta_1}{\theta_2}), \qquad \mu_2 = \frac{\theta_1(\theta_3 - \tilde{\theta}_2)}{\theta_2(\theta_3 - \theta_1)} \in (0, \kappa), \tag{6.26}$$

where $\kappa = \theta_1(\theta_3 - \theta_2)/(\theta_2(\theta_3 - \theta_1))$. Recasting of α_0^2, β_0^2, α_1^2 and β_1^2 yields

$$\alpha_0^2 = \frac{(\frac{\theta_1}{\theta_2} - \mu_1)(\frac{\theta_1}{\theta_2} - \mu_2)\theta_2}{\kappa(\mu_1 - \kappa)(\kappa - \mu_2)\theta_1}, \qquad \beta_0^2 = \frac{(\frac{\theta_1}{\theta_2} - \kappa)(\frac{\theta_1}{\theta_2} - \mu_1)(\frac{\theta_1}{\theta_2} - \mu_2)\theta_3^2}{\kappa\mu_1\mu_2\theta_1^2}, \tag{6.27}$$
$$\alpha_1^2 = \frac{\theta_1(\frac{\theta_1}{\theta_2} - \mu_1)(\kappa - \mu_2)}{\theta_2\kappa(\mu_1 - \kappa)(\frac{\theta_1}{\theta_2} - \mu_2)}, \qquad \beta_1^2 = \frac{(\frac{\theta_1}{\theta_2} - \kappa)(\frac{\theta_1}{\theta_2} - \mu_1)\mu_2}{\kappa\mu_1(\frac{\theta_1}{\theta_2} - \mu_2)}. \tag{6.28}$$

We collect the results to represent

$$\frac{\Delta_{1,2}^{[-3]}(\lambda')}{\Delta_{1,2}^{[-3]}(\lambda)} = \frac{\tilde{\theta}_1 - \theta_1}{\theta_2 - \tilde{\theta}_1} \cdot \frac{\dfrac{\theta_2 - \theta_1}{\theta_3 - \theta_1} - \beta_0^2 \dfrac{(\theta_3 - \theta_2)\theta_1}{(\theta_3 - \theta_1)\theta_2}}{\dfrac{\theta_2 - \theta_1}{\theta_3 - \theta_1}\alpha_0^2 + \beta_0^2}.$$

In order to eliminate $\tilde{\theta}_1$, $\tilde{\theta}_2$ and θ_3 check that

$$\frac{\theta_2-\theta_1}{\theta_3-\theta_1}=1-\frac{\theta_2}{\theta_1}\kappa, \qquad \frac{\tilde{\theta}_1-\theta_1}{\theta_3-\theta_1}=1-\frac{\theta_2}{\theta_1}\mu_1, \qquad \frac{\theta_2-\tilde{\theta}_1}{\theta_3-\theta_1}=\frac{\theta_2}{\theta_1}(\mu_1-\kappa),$$

so that

$$\frac{\Delta_{1,2}^{[-3]}(\lambda')}{\Delta_{1,2}^{[-3]}(\lambda)}=\frac{\frac{\theta_1}{\theta_2}-\mu_1}{\mu_1-\kappa}\cdot\frac{\frac{\theta_1}{\theta_2}-\kappa-\beta_0^2\kappa}{\left(\frac{\theta_1}{\theta_2}-\kappa\right)\alpha_0^2+\frac{\theta_1}{\theta_2}\beta_0^2}.$$

Inserting (6.27) together with $\omega=\theta_1/\theta_2$ results in

$$\frac{\Delta_{1,2}^{[-3]}(\lambda')}{\Delta_{1,2}^{[-3]}(\lambda)}=\frac{\kappa(\kappa-\mu_2)}{(\omega-\mu_2)}\cdot\frac{\mu_1\mu_2-(\omega-\mu_1)(\omega-\mu_2)\theta_3^2\theta_1^{-2}}{\omega^{-1}\mu_1\mu_2+\omega\theta_3^2\theta_1^{-2}(\mu_1-\kappa)(\kappa-\mu_2)}=:h(\mu_1,\mu_2).$$

Since $h(\cdot,\mu_2)$ in a monotone increasing function in $\mu_1\in(\kappa,\omega)$ we find the reduced representation

$$\sup_{\mu_1\in(\kappa,\omega)}h(\mu_1,\mu_2)=h(\omega,\mu_2)=\frac{\kappa(\kappa-\mu_2)}{\omega-\mu_2}\cdot\frac{\omega\mu_2}{\mu_2+\omega^2\frac{\theta_3^2(\theta_2-\theta_1)}{\theta_1^2(\theta_3-\theta_1)}(\kappa-\mu_2)}=:g(\mu_2).$$

The maximum of $g(\mu_2)$ is taken in

$$\mu_2^*=\kappa\rho\omega\frac{1-\theta_2/\theta_3}{\omega(\rho-1)+\kappa}$$

with

$$\rho=\frac{\theta_3^2(\theta_2-\theta_1)}{\theta_2^2(\theta_3-\theta_1)}.$$

The proof of (6.21) is completed by tedious simplifications leading to

$$g(\mu_2^*)=\left(\frac{\kappa}{2-\kappa}\right)^2.$$

In order to show (6.14), first check that under the assumption $\|x_1\|=1$ together with (6.24) it holds

$$\tan^2\varphi_0=\frac{1-((x,x_1)/\|x\|)^2}{((x,x_1)/\|x\|)^2}=\alpha_0^2+\beta_0^2.$$

To express $(\tan^2\varphi_1)/(\tan^2\varphi_0)$ in terms of μ_i we plug in (6.26) and find

$$\begin{aligned}\frac{\tan^2\varphi_1}{\tan^2\varphi_0}&=\frac{\alpha_1^2+\beta_1^2}{\alpha_0^2+\beta_0^2}\\&=\frac{(\kappa-\mu_2)\theta_1^2\mu_2}{(\frac{\theta_1}{\theta_2}-\mu_2)^2\theta_2}\cdot\frac{\theta_1\mu_1(\kappa-\mu_2)+\theta_2\mu_2(\frac{\theta_1}{\theta_2}-\kappa)(\mu_1-\kappa)}{\theta_1\theta_2\mu_1\mu_2+\theta_3^2(\frac{\theta_1}{\theta_2}-\kappa)(\mu_1-\kappa)(\kappa-\mu_2)}=:g(\mu_1,\mu_2).\end{aligned}$$

Partial differentiation of $g(\mu_1, \mu_2)$ with respect to μ_1 shows that

$$\operatorname{sgn}\left(\frac{\partial g}{\partial \mu_1}\right) = \operatorname{sgn}\left(\mu_2(1+\frac{\theta_2}{\theta_3}) - \kappa\right).$$

We first scrutinize the case of a monotone decreasing $g(\mu_1, \mu_2)$ for $\mu_1 \in (\kappa, \frac{\theta_1}{\theta_2})$ and $0 < \mu_2 \leq \frac{\kappa}{1+\theta_2/\theta_3}$. Then

$$\sup_{\mu_1 \in (\kappa, \frac{\theta_1}{\theta_2})} g(\mu_1, \mu_2) = g(\kappa, \mu_2) = \frac{(\kappa-\mu_2)^2\theta_1^2}{(\theta_1 - \mu_2\theta_2)^2} =: h(\mu_2).$$

Since

$$\frac{\partial}{\partial \mu_2} h(\mu_2) = \frac{2\theta_1^2(\mu_2-\kappa)(\frac{\theta_1}{\theta_2}-\kappa)}{\theta_2^2\left(\frac{\theta_1}{\theta_2}-\mu_2\right)^3} < 0,$$

one is led to

$$\sup_{\mu_2 \in (0, \frac{\kappa}{1+\theta_2/\theta_3})} h(\mu_2) = h(0) = \kappa^2.$$

In the second case $\kappa/(1+\theta_2/\theta_3) \leq \mu_2 \leq \kappa$ it holds

$$\sup_{\mu_1 \in (\kappa, \frac{\theta_1}{\theta_2})} g(\mu_1, \mu_2) = g(\frac{\theta_1}{\theta_2}, \mu_2) = \frac{(\kappa-\mu_2)\theta_1^2\mu_2}{(\frac{\theta_1}{\theta_2}-\mu_2)^2\theta_2} \cdot \frac{\theta_1^2(\kappa-\mu_2) + \mu_2\theta_2^2\left(\frac{\theta_1}{\theta_2}-\kappa\right)^2}{\theta_1^2\mu_2 + \theta_3^2(\kappa-\mu_2)\left(\frac{\theta_1}{\theta_2}-\kappa\right)^2} < \kappa^2,$$

where the last inequality results from strenuous direct calculations. □

The proof of the following corollary can be done by transferring the arguments of Knyazev and Skorokhodov [76] to the actual setup. We note that the estimate (6.14) is only attained asymptotically for $\lambda_n \to \infty$. (Similarly, the correct form of Equation (2.4) in [76] has to be equipped with an additional limit $\lambda_1 \to \infty$.)

Corollary 6.9. *The estimate (6.13) is asymptotically sharp in the sense that*

$$\lim_{\epsilon \to 0} \sup_{\measuredangle(x,x_1)\leq\epsilon} \frac{\Delta_{1,2}(\lambda(x^{(k)}))}{\Delta_{1,2}(\lambda(x^{(0)}))} = \left(\frac{1-\xi}{1+\xi}\right)^{2k}.$$

Furthermore, it holds

$$\lim_{\lambda_n \to \infty} \sup_{x^{(k)} \neq 0} \frac{\tan^2 \varphi_{k+1}}{\tan^2 \varphi_k} = (1-\xi)^2.$$

6.3 PINVIT(2) convergence theory

Our convergence analysis of PINVIT(2) consists of the following steps: In Section 6.3.1 some elementary results on the fastest/poorest convergence are collected. Then the c-basis formulation of PINVIT(2) is introduced in Section 6.3.2, setting up the appropriate geometry as described in Section 6.3.3. Finally, in Section 6.3.4 a conjecture on the 3D subspace of poorest convergence is given, which paves the way for the mini-dimensional analysis in the subsequent Section 6.4. Finally, Section 6.5 reports on the results of several numerical experiments supporting the validity of the conjecture on poorest convergence in a low dimensional invariant subspace.

6.3.1 Elementary results on extremal convergence

Let us collect some results on the fastest convergence of PINVIT(2) as well as its poorest convergence in the domain of Rayleigh quotients $[\lambda_{n-1}, \lambda_n]$. Fortunately, the best possible convergence of PINVIT(2) is considerably easier to analyze than that of PINVIT(1). Lemma 6.10 reveals that for any $\lambda \in [\lambda_1, \lambda_n)$ some vector x in the level set

$$L(\lambda) = \{x \in \mathbb{R} : \ \lambda(x) = \lambda\} \tag{6.29}$$

together with a preconditioner can be specified, in such a way that PINVIT(2) applied to x terminates immediately within the smallest eigenpair (λ_1, x_1). The comparison with the results on fastest PINVIT(1) convergence, as gained in Chapter 3, highlights this superior convergence property of PINVIT(2). Furthermore, the fastest convergence of PINVIT(2) does not depend on γ, in contrast to PINVIT, which may converge faster for increasing γ.

Lemma 6.10 (Fastest convergence of PINVIT(2)). *Let $\lambda \in [\lambda_1, \lambda_n)$. Then*

$$\lambda_1 = \min_{x \in L(\lambda)} \min_{B^{-1} \in \mathcal{B}_\gamma} \min_{\omega \in \mathbb{R}} \lambda(x - \omega B^{-1}(Ax - \lambda x)),$$

with $L(\lambda)$ defined by (6.29) and where the set $\mathcal{B}_\gamma$, see (2.3), contains all admissible preconditioners for some $\gamma \in [0, 1)$.

Proof. Define $x \in \operatorname{span}\{x_1, x_n\}$ by

$$x = \left(\left(\frac{\lambda_n - \lambda}{\lambda_n - \lambda_1} \right)^{1/2}, 0, \cdots, 0, \left(\frac{\lambda - \lambda_1}{\lambda_n - \lambda_1} \right)^{1/2} \right), \tag{6.30}$$

so that $x \in L(\lambda)$. Since for any $\gamma \in [0, 1)$ the center $\lambda A^{-1}x$ is contained in $E_\gamma(x)$, the smallest Ritz value with respect to

$$\operatorname{span}\{x, \lambda A^{-1}x\} = \operatorname{span}\{x_1, x_n\}$$

is given by λ_1. □

Lemma 6.10 indicates that the vector of poorest convergence of PINVIT(2), aside from $\lambda \in (\lambda_n, \lambda_{n-1})$, is spanned by at least three eigenvectors. Therefore, let $Z_{i,j} \in \mathcal{B}_\gamma$ be an operator having $\mathrm{span}\{x_i, x_j\}$ as an invariant subspace. With this choice we cover the case of exact preconditioning, i.e. $Z_{i,j} = A^{-1}$, as well as all preconditioners responsible for the poorest convergence of PINVIT(1) in $\mathrm{span}\{x_i, x_j\}$ for $j = i + 1$.

Corollary 6.11 (Subspace of poorest convergence). *Let $\lambda \in [\lambda_1, \lambda_{n-1})$, $\lambda \neq \lambda_i$ and let*

$$x^* \in \arg \max_{x \in L(\lambda)} \min_{\omega \in \mathbb{R}} \lambda(x - \omega Z_{i,j}(Ax - \lambda x)). \tag{6.31}$$

Then $x^* \notin \mathrm{span}\{x_i, x_j\}$ for any $1 \leq i, j \leq n$.

Proof. Assume $x \in \mathrm{span}\{x_i, x_j\}$ with $i < j$. Then both $Ax - \lambda x$ and $Z_{i,j}(Ax - \lambda x)$ are contained in $\mathrm{span}\{x_i, x_j\}$. Consequently, PINVIT(2) would immediately terminate in the eigenpair (x_i, λ_i). It is easy but wearisome to construct examples in $L(\lambda)$ so that for exact preconditioning with $A^{-1} \in \mathcal{B}_0 \subset \mathcal{B}_\gamma$ it holds that

$$\lambda_i < \max_{x \in L(\lambda)} \min_{\omega \in \mathbb{R}} \lambda(x'(\omega)).$$

□

We conclude from Corollary 6.11 that poorest convergence of INVIT(2) is taken at least in a 3D space as has also been suggested by Theorem 6.3. Note that Theorem 6.3 does not provide a proof of this 3D property since the bound (6.13) is only attainable for $\lambda \to \lambda_1$. For PINVIT(2) Corollary 6.11 has two possible consequences: either poorest convergence is at least taken in a 3D space, or the preconditioner responsible for the poorest convergence does not have $\mathrm{span}\{x_i, x_j\}$ as an invariant subspace.

Poorest convergence of PINVIT(2) in the domain $[\lambda_{n-1}, \lambda_n)$ is treated in Lemma 6.12 by using the Courant-Fischer theorem.

Lemma 6.12 (Poorest convergence for $\lambda \geq \lambda_{\mathbf{n-1}}$). *Let $\lambda \in [\lambda_{n-1}, \lambda_n)$ and x^* be the vector of poorest PINVIT(2) convergence w.r.t. $L(\lambda)$ and $\mathcal{B}_\gamma$. Then $x^* \in \mathrm{span}\{x_{n-1}, x_n\}$ and*

$$\lambda_{n-1} = \max_{x \in L(\lambda)} \max_{B^{-1} \in \mathcal{B}_\gamma} \min_{\omega \in \mathbb{R}} \lambda(x - \omega B^{-1}(Ax - \lambda)x).$$

Proof. Let $\theta_1 \leq \theta_2$ be the Ritz values defined by $\mathrm{span}\{x, B^{-1}(Ax - \lambda x)\}$. By the Courant-Fischer criteria it holds that $\theta_1 \leq \lambda_{n-1}$. The last inequality is attained for $x \in \mathrm{span}\{x_{n-1}, x_n\}$ and $B = A$. □

Let us summarize that the fastest convergence of PINVIT(2) is taken in 2D invariant subspaces of A while its poorest convergence (at least for $\gamma = 0$) is attained in at least 3D invariant subspaces. This should be seen in contrast to PINVIT(1), for which the best and poorest convergence are both taken in (differing) 2D invariant subspaces.

6.3.2 The c-basis representation

For the remaining part of this chapter we prefer to analyze PINVIT(2) within the c-basis which expresses the underlying geometry in an advantageous way and which has already been approved as a valuable tool for the analysis of PINVIT(1). We first recall the normal form of our symmetric positive definite preconditioners

$$B^{-1} = A^{-1} + A^{-1/2}\hat{V}D\hat{V}^T A^{-1/2},$$

cf. Equation (2.8), for arbitrary orthogonal matrices $\hat{V}$ and diagonal matrices D with $d_{ii} \in [-\gamma, \gamma]$ so that the spectral radius of $I - B^{-1}A$ is bounded by γ. By inserting these preconditioners into PINVIT(2) and applying the c-basis transformation (2.17) one obtains for the search subspace

$$\begin{aligned} V &= [x, B^{-1}(Ax - \lambda x)] \\ &= [X\Lambda^{-1/2}c, X(X^TA^{-1} + X^TA^{-1/2}X(X^T\hat{V})D(X^T\hat{V})^TX^TA^{-1/2})(A - \lambda I)X\Lambda^{-1/2}c] \\ &= A^{-1/2}X\ [c, (I + UDU^T)(c - \lambda\Lambda^{-1/2}c)], \end{aligned}$$

with the orthogonal matrix $U = X^T\hat{V}$. Then PINVIT(2) reads

$$c' = c - \omega(I + UDU^T)(c - \lambda\Lambda^{-1}c), \tag{6.32}$$

where ω is computed so that the Rayleigh quotient $(c', c')/(c', \Lambda^{-1}c')$ is minimized. In terms of the c-basis (6.3) and (6.4) read

$$\bar{A} = [c, (I + UDU^T)(c - \lambda\Lambda^{-1/2}c)]^T[c, (I + UDU^T)(c - \lambda\Lambda^{-1/2}c)], \tag{6.33}$$

$$\bar{M} = [c, (I + UDU^T)(c - \lambda\Lambda^{-1/2}c)]^T\Lambda^{-1}[c, (I + UDU^T)(c - \lambda\Lambda^{-1/2}c)]. \tag{6.34}$$

In order to show that the convergence analysis can be restricted to nonnegative vectors c let us introduce the sign-changing operator

$$P = \mathrm{diag}(\sigma_1, \ldots, \sigma_n), \qquad \sigma_i \in \{1, -1\}.$$

One benefit of our geometrical description of PINVIT(2) comes to light in Lemma 6.13. While it is complicated to understand how PINVIT(2) behaves for a *fixed* preconditioner as c is replaced by Pc, its effect on $E_\gamma(c)$ and $E_\gamma(Pc)$ can be understood by elementary geometric arguments. Lemma 6.13 reveals that such a replacement has no effect on the set of Rayleigh quotients that can be attained by PINVIT(2).

Lemma 6.13. *For given nonzero c and $\gamma \in [0, 1)$ let $\Sigma(c)$ be the set of PINVIT(2)-attainable Rayleigh quotients for all admissible preconditioners*

$$\Sigma(c) = \{\lambda(c') \text{ by Equation (6.32)}\ :\ D \text{ with } |d_{ii}| \le \gamma,\ \ \text{orthogonal } U \in \mathbb{R}^{n\times n}\}.$$

Then

$$\Sigma(c) = \Sigma(Pc).$$

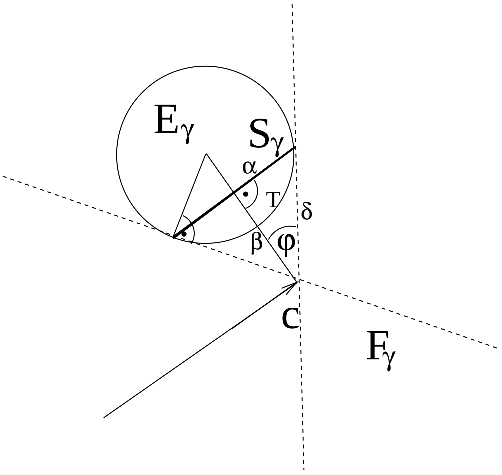


Figure 6.2: *Cross section along* $\operatorname{span}\{c, \lambda\Lambda^{-1}c\}$ *through* $E_\gamma(c)$*, the cone* F_γ *and* S_γ*.*

Proof. As shown in Section 2.2, changing the sign of the ith component of c acts like a reflection of $E_\gamma(c)$ through the hyperplane $\operatorname{span}\{e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_n\}$ so that

$$PE_\gamma(c) = E_\gamma(Pc).$$

For any $y \in E_\gamma(c)$ we have

$$\begin{aligned} \lambda' &= \min_\omega \lambda(y + \omega c) = \min_\omega \frac{(y, y) + 2\omega(y, c) + \omega^2(c, c)}{(y, \Lambda^{-1}y) + 2\omega(y, \Lambda^{-1}c) + \omega^2(c, \Lambda^{-1}c)} \\ &= \frac{(Py, Py) + 2\omega(Py, Pc) + \omega^2(Pc, Pc)}{(Py, \Lambda^{-1}Py) + 2\omega(Py, \Lambda^{-1}Pc) + \omega^2(Pc, \Lambda^{-1}Pc)} \\ &= \min_\omega \lambda(Py + \omega Pc), \end{aligned}$$

since $(c, y) = \sum_{i=1}^n \sigma_i^2 c_i y_i = (Pc, Py)$. We infer that in $\operatorname{span}\{y, c\}$ and $\operatorname{span}\{Py, Pc\}$ the same extremal Rayleigh quotients, or Ritz values, are taken. □

6.3.3 A geometric representation

In order to work out a geometric picture of PINVIT(2) let us consider the scheme

$$c'(\omega) = c - \omega(I + UDU^T)(c - \lambda\Lambda^{-1}c) \tag{6.35}$$

for all $\omega \in \mathbb{R}$. Figure 6.2 in $\operatorname{span}\{c, \lambda\Lambda^{-1}c\}$ displays the set $E_\gamma(c)$ of admissible vectors thrown out by PINVIT for a fixed $\gamma \in [0, 1)$ as well as the (dashed) cone

$$F_\gamma(c) = \{c'(\omega) : \ \omega \in \mathbb{R}, \ \text{orthogonal } U \in \mathbb{R}^{n\times n}, |d_{ii}| \in [0, \gamma]\}.$$

$F_\gamma(c)$ is the smallest circular cone containing $E_\gamma(c)$ and having its vertex in c. In other words, $F_\gamma(c)$ includes all the lines $c - \omega y$, $\omega \in \mathbb{R}$, where $y = (I + UDU^T)(c - \lambda\Lambda^{-1}c)$ is some search direction. PINVIT(2) minimizes the Rayleigh quotient along each of these lines. It is easy to show by numerical examples that one cannot restrict the analysis to the half-cone $\omega \geq 0$, because minima of the Rayleigh quotient on these lines through c may surprisingly be taken in each of the half spaces defined by the tangent manifold of $E_1(c)$ in c.

By our geometric construction the set of all possible search directions is uniquely determined by $S_\gamma(c) \subseteq E_\gamma(c)$ which is given by

$$S_\gamma(c) := \{c + (1-\gamma^2)(\lambda\Lambda^{-1}c - c) + \alpha v : \ \|v\| \leq 1, v \perp c - \lambda\Lambda^{-1}c\}, \tag{6.36}$$

where φ with $\sin\varphi = \gamma$ denotes the opening angle of $F_\gamma(c)$ and the quantities α, β and δ are the sides of the smaller right triangle T shown in Figure 6.2 with

$$\alpha = \gamma(1-\gamma^2)^{1/2}\|c - \lambda\Lambda^{-1}c\|, \tag{6.37}$$

$$\beta = (1-\gamma^2)\|c - \lambda\Lambda^{-1}c\|, \tag{6.38}$$

$$\delta = (1-\gamma^2)^{1/2}\|c - \lambda\Lambda^{-1}c\|. \tag{6.39}$$

In order to clearify the assignment of α, β and γ we repeat the definition of the opening angle

$$\sin\varphi = \frac{\alpha}{\delta} = \gamma.$$

PINVIT(2) is adequately described (in a sense that any admissible search direction is contained in $S_\gamma(c) - c$) as a mapping on $S_\gamma(c) - c$

$$\Pi_0 : (S_\gamma(c) - c) \to P_\gamma(c) : y \mapsto c + \left[\arg\min_\omega(\lambda(c + \omega y))\right] y.$$

Unfortunately, this preliminary geometric picture of PINVIT(2) is handicapped by the fact that $P_\gamma(c)$ is an unbounded set if γ exceeds some critical value γ_l.

Lemma 6.14. *Whenever* $c_1 \neq 0$ *and*

$$\gamma > \gamma_L := \frac{\|c - \lambda\Lambda^{-1}c - e_1(e_1, c - \lambda\Lambda^{-1}c)\|}{\|c - \lambda\Lambda^{-1}c\|},$$

then Π_0 *on* $S_\gamma(c) - c$ *has a discontinuity in* e_1 *yielding an unbounded set* $P_\gamma(c)$.

Proof. Since $c_1 \neq 0$ we have $(\lambda\Lambda^{-1}c - c, e_1) \neq 0$. Therefore, $c + e_1$ is not in the plane tangential to $E_1(c)$ in c and we determine the smallest γ, called γ_L, so that $c + e_1 \in F_\gamma(c)$. Since $\|c + \vartheta e_1 - \lambda\Lambda^{-1}c\|$ is minimized in $\vartheta = (e_1, \lambda\Lambda^{-1}c - c)$ we obtain for γ_L

$$\gamma_L\|c - \lambda\Lambda^{-1}c\| = \|c - \lambda\Lambda^{-1}c - e_1(e_1, c - \lambda\Lambda^{-1}c)\|.$$

Finally, note that $\lim_{\omega\to\pm\infty}\lambda(c + \omega e_1) = \lambda_1$ is responsible for the singularity. □

In order to avoid this discontinuity we analyze instead of (6.35)

$$c'(\vartheta^*) = \vartheta^* c - (I + UDU^T)(c - \lambda\Lambda^{-1}c) \tag{6.40}$$

where ϑ^* is computed so that the Rayleigh quotient of c' is minimized. The benefit of this alternative scaling strategy is elucidated in the next lemma.

Lemma 6.15. *If $0 < \gamma < 1$, then ϑ^*, as defined in (6.40), is bounded and so is $Q_\gamma(c)$, the image of $S_\gamma(c) - c$ under our modified representation (6.40) of PINVIT(2).*

$$\Pi : (S_\gamma(c) - c) \to Q_\gamma(c) : y \mapsto \left[\arg\min_{\vartheta} \lambda(\vartheta c - y)\right] c - y. \tag{6.41}$$

The mapping Π defines the new PINVIT(2) iterate for any search direction from $S_\gamma(c) - c$ in a unique way (i.e. the scaling of $\Pi(y)$ is immaterial).

Proof. Obviously, for any search direction $y \in S_\gamma(c) - c$ it holds

$$\lim_{|\vartheta|\to\infty} \lambda(\vartheta c - y) = \lambda(c),$$

where λ is a continuous function in ϑ. As PINVIT(2) decreases the Rayleigh quotient more rapidly than PINVIT(1), we have $\lambda(\vartheta c - y) \le \lambda(c - y) < \lambda(c)$, which entails boundedness of ϑ. (If y is collinear to e_1, then ϑ may equal 0.) □

6.3.4 A conjecture on the subspace of poorest convergence

As a result of Corollary 6.11, the space in which INVIT(2) takes its poorest convergence is spanned by at least 3 eigenfunctions of A. We have not succeeded in proving that poorest convergence of PINVIT(2) is taken in the 3D space $\mathrm{span}\{e_i, e_{i+1}, e_n\}$ corresponding to λ_i, λ_{i+1} and λ_n as formulated in Conjecture 6.16.

Conjecture 6.16. *Let $L(\lambda) = \{c \in \mathbb{R}^n : \lambda(c) = \lambda\}$, $\lambda \in (\lambda_i, \lambda_{i+1})$, $i + 1 < n$, be the level set of the Rayleigh quotient. Then*

$$\arg \sup_{c\in L(\lambda)} \sup_{U^TU=I} \sup_{|d_{ii}|\in[0,\gamma]} \inf_{\vartheta\in\mathbb{R}} \lambda(\vartheta c - (I + UDU^T)(c - \lambda\Lambda^{-1}c)) \in \mathrm{span}\{e_i, e_{i+1}, e_n\},$$

where the supremum in U is taken over all orthogonal matrices $U \in \mathbb{R}^{n\times n}$ and where d_{ii}, $1 \le i \le n$, denote the diagonal elements of the diagonal matrix D.

There are some good reasons which make Conjecture 6.16 appear very reasonable. On the one hand, the mini-dimensional analysis of INVIT(2) (limit case of PINVIT(2) for $\gamma = 0$) in Section 6.2 is given within the 3D subspace $H^{[-3]}$ having the Ritz values λ_1, θ_2 and θ_3. We get an asymptotically sharp estimate from (6.13) if $\theta_2 = \lambda_2$ and $\theta_3 = \lambda_n$ for some specific

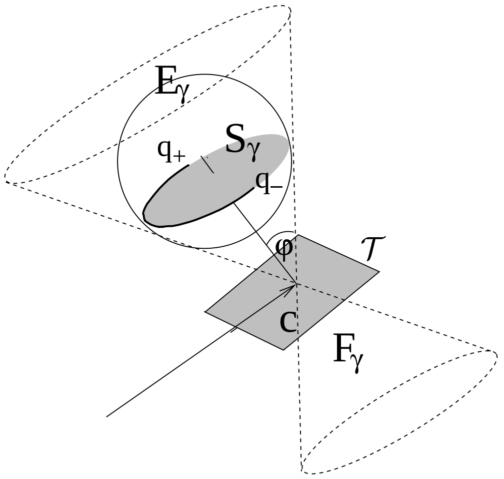


Figure 6.3: *Poorest convergence of PINVIT(2) in* $\operatorname{span}\{e_1, e_2, e_n\}$.

choice of the iteration vector. Additionally, be aware of the fact that such a property of mini-dimensionality is well known from the PINVIT analysis where (if $\lambda \in (\lambda_i, \lambda_{i+1})$)

$$\arg \sup_{c \in L(\lambda)} \sup_{U^T U = I} \sup_{|d_{ii}| \in [0,\gamma]} \lambda(c - (I + UDU^T)(c - \lambda\Lambda^{-1}c)) \in \operatorname{span}\{e_i, e_{i+1}\}.$$

On the other hand, there is a clear numerical evidence confirming Conjecture 6.16. Numerical data are presented in Section 6.5.

6.4 Mini-dimensional analysis

Supposing that Conjecture 6.16 holds, it suffices to analyze PINVIT(2) within a three-dimensional space in order to derive convergence estimates. Hence we assume in the following $c \in \mathbb{R}^3$; later we apply the results to 3D subspace $\operatorname{span}\{e_i, e_{i+1}, e_n\}$; for simplicity we sometimes write $i = 1$. The present geometry in $\mathbb{R}^3$ can be looked up in Figure 6.3. By Equation (6.36) it holds that

$$S_\gamma(c) - (1 - \gamma^2)(\lambda\Lambda^{-1}c - c) \subseteq \mathcal{T}, \tag{6.42}$$

where $\mathcal{T}$ denotes the tangential plane of $E_1(c)$ in c. In other words S_γ and $\mathcal{T}$ are parallel planes since any $z \in S_\gamma(c) - (1-\gamma^2)(\lambda\Lambda^{-1}c - c)$ has the form $c + \xi\alpha v$, α by (6.37), $-1 \le \xi \le 1$ and $v \perp (I - \lambda\Lambda^{-1})c$ so that

$$(c + \xi\alpha v, (I - \lambda\Lambda^{-1})c) = 0.$$

Applying PINVIT(2) in the form (6.41) provides the justification to remove some degree of freedom from the set of search directions $S_\gamma(c)$ as elucidated in Lemma 6.17.

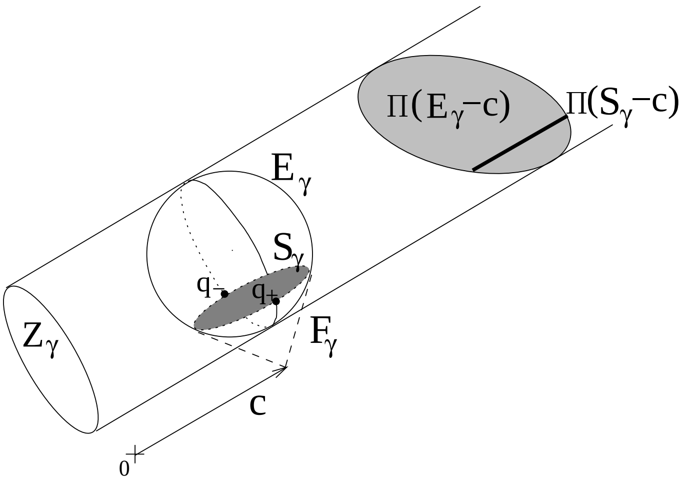


Figure 6.4: *Construction of* q_+ *and* q_-.

Lemma 6.17. *PINVIT(2) as defined in Lemma 6.15 without loss of generality can be restricted to the 1D curve* $\partial S_\gamma(c)$ *with* $\partial S_\gamma(c) := \partial E_\gamma(c) \cap S_\gamma(c)$ *and*

$$\Pi : (\partial S_\gamma(c) - c) \to Q_\gamma(c) : y \mapsto \left[\arg\min_\vartheta(\lambda(\vartheta c - y))\right] c - y$$

remains to be a surjective mapping $Q_\gamma(c)$.

Proof. By Equation (6.42) any $y \in \mathring{S}_\gamma(c) := \mathring{E}_\gamma(c) \cap S_\gamma(c)$ can be written as $y = \tilde{y} + c$ with $\tilde{y} \in \partial S_\gamma(c)$. Thus,

$$\begin{aligned} \Pi(y) &= \left[\arg\min_\vartheta \lambda(\vartheta c - y)\right] c - y = \vartheta^* c - y = (\vartheta^* + 1)c - (y - c) \\ &= \left[\arg\min_\vartheta \lambda(\vartheta c - \tilde{y})\right] c - \tilde{y} = \Pi(\tilde{y}). \end{aligned}$$

□

Obviously, one can go beyond the result of Lemma 6.17 in such a manner that there is a half-circle as a minimal subset of $\partial S_\gamma(c)$ so that any two points of this half-circle are not connected by a multiple of c. Even on this half-circle Π is surjective. Theorem 6.18 discloses that only the two end points of the half-circles mentioned above are responsible for the poorest convergence.

Theorem 6.18. *As long as preconditioned steepest descent on the set of search directions* S_γ *does not terminate within the smallest eigenpair* (e_1, λ_1), *it is of poorest convergence in* q_+ *or* q_- *with* $q_\pm \in S_\gamma(c)$ *defined by*

$$q_\pm = c + (1-\gamma^2)(\lambda\Lambda^{-1}c - c) \pm \gamma(1-\gamma^2)^{1/2} \|(I - \lambda\Lambda^{-1})c\| \, \frac{c \times \Lambda^{-1}c}{\|c \times \Lambda^{-1}c\|}. \tag{6.43}$$

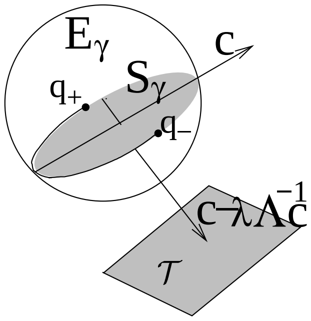


Figure 6.5: *Points q_+ and q_- of extremal convergence.*

Proof. We start with the definition of the smallest circular infinite cylinder $Z_\gamma(c)$ enclosing $E_\gamma(c)$ and also containing $E_\gamma(c)+\rho c$, $\rho\in\mathbb{R}$, i.e.

$$Z_\gamma(c) := \{z+\rho c:\ z\in E_\gamma(c),\ \rho\in\mathbb{R}\},$$

see Figure 6.4. Since $S_\gamma(c)+(1-\gamma)\gamma(c-\lambda\Lambda^{-1}c)$ is a subset of the tangential plane of $Z_\gamma(c)$ in $\lambda\Lambda^{-1}c+\gamma(I-\lambda\Lambda^{-1})c$ there are exactly two elements in $S_\gamma(c)\cap Z_\gamma(c)$ for $\gamma\in(0,1)$. By using (6.37)–(6.39), these two points called q_+ and q_- have the form (6.43). Next we show that no search direction from the interior of $E_\gamma(c)$ embodies the poorest convergence. Let $c+d\in\mathring{E}_\gamma(c)$ and let its image under Π be given by $\vartheta^* c+d$. Then

$$\frac{d}{d\vartheta^*}\lambda(\vartheta^* c+d) = (\nabla\lambda(\vartheta^* c+d),c) = 0.$$

If $w=\nabla\lambda(\vartheta^* c+d)=0$, then $\vartheta^* c+d$ would be an eigenvector where collinearity to e_1 is excluded by the assumption. The remaining eigenvectors e_i, $2\le i\le n-1$ can be excluded by an analysis of the Hessian of $\lambda(\cdot)$ since all these eigenvectors are saddle points of the Rayleigh quotient. The image of $\Pi(E_\gamma(c)-c)$ is a 2D continuously differentiable manifold $\mathcal{M}$ (ϑ^* is a differentiable function of y). Now, $c\perp w$ implies that the projection $P_{\mathcal{M}}w$ to the tangential plane of $\Pi(E_\gamma(c)-c)$ in $\vartheta^* c+d$ does not vanish. (Note that Lemma 6.17 shows that Π is essentially a mapping from any 2D cross section through $Z_\gamma(c)$ on $\mathcal{M}$.) Therefore, in a neighborhood of $\vartheta^* c+d$ another point $\tilde\vartheta c+\tilde d$ with an increased Rayleigh quotient can be found, which is the image of some $c+\tilde d\in\mathring{E}_\gamma(c)$. □

Remark 6.19. *The proof of Theorem 6.18 provides additional information concerning the search directions of poorest convergence for the general n-dimensional case. All the arguments showing that these directions are localized on $S_\gamma(c)\cap Z_\gamma(c)$ even hold in the $\mathbb{R}^n$ while only the part of the construction of $q_\pm$ is restricted to $\mathrm{span}\{e_1,e_2,e_n\}$. In Section 6.5 we make use of this property in order to give some numerical evidence for Conjecture 6.16.*

Since $q_\pm$ by Theorem 6.18 are the only possible candidates for the poorest convergence we write down the associated admissible (and appropriately scaled) search directions of PINVIT(2)

$$d_\pm := (1-\gamma^2)^{1/2}\frac{\lambda\Lambda^{-1}c - c}{\|(I-\lambda\Lambda^{-1})c\|} \pm \gamma\frac{c\times\Lambda^{-1}c}{\|c\times\Lambda^{-1}c\|}.$$

Our next aim is to find out which of the search directions $d_\pm$ is responsible for the poorest convergence of PINVIT(2). Therefore, we apply the Rayleigh-Ritz procedure (see Equations (6.33) and (6.34) for its c-basis formulation) to the two subspaces $[c, d_+]$ and $[c, d_-]$. We determine in each of these subspaces the smallest Ritz values corresponding to each of the new PINVIT(2) iterates. The larger one of these Ritz values will present the case of poorest convergence. Next we determine the projection matrices.

We get the pleasant result $\bar{A} = I \in \mathbb{R}^{2\times 2}$, since $\|c\| = 1$ and

$$\begin{aligned}(c, d_\pm) &= (c, (1-\gamma^2)^{1/2}\frac{\lambda\Lambda^{-1}c - c}{\|(I-\lambda\Lambda^{-1})c\|} \pm \frac{c\times\Lambda^{-1}c}{\|c\times\Lambda^{-1}c\|}) = 0,\\ (d_\pm, d_\pm) &= (1-\gamma^2)\frac{(\lambda\Lambda^{-1}c - c, \lambda\Lambda^{-1}c - c)}{\|(I-\lambda\Lambda^{-1})c\|^2} + \gamma^2\|v\|^2 = 1.\end{aligned}$$

Therefore, solving the Rayleigh-Ritz generalized eigenvalue problem $(\bar{A}, \bar{M})$ equals computing the inverse eigenvalues of $\bar{M} = [c, d_\pm]^T\Lambda^{-1}[c, d_\pm]$. One derives

$$\bar{m}_{11} = 1/\lambda,$$

and

$$\begin{aligned}\bar{m}_{12} = \bar{m}_{21} &= \frac{(1-\gamma^2)^{1/2}}{\|(I-\lambda\Lambda^{-1})c\|}(\lambda\Lambda^{-1}c - c, \Lambda^{-1}c) \pm \gamma(\frac{c\times\Lambda^{-1}c}{\|c\times\Lambda^{-1}c\|}, \Lambda^{-1}c)\\ &= \frac{(1-\gamma^2)^{1/2}}{\lambda\|(I-\lambda\Lambda^{-1})c\|}(\lambda\Lambda^{-1}c - c, \lambda\Lambda^{-1}c)\\ &= \frac{(1-\gamma^2)^{1/2}}{\lambda}\|(I-\lambda\Lambda^{-1})c\|.\end{aligned}$$

Hence $\bar{m}_{11}$ and $\bar{m}_{12}$ are the same for each d_+ and d_-. Finally, we have

$$\begin{aligned}\bar{m}_{22} = (1-\gamma^2)&\frac{((I-\lambda\Lambda^{-1})c, \Lambda^{-1}((I-\lambda\Lambda^{-1})c))}{\|(I-\lambda\Lambda^{-1})c\|^2}\\ &\pm \frac{\lambda(1-\gamma^2)^{1/2}\gamma}{\|(I-\lambda\Lambda^{-1})c\|}(\Lambda^{-2}c, v) + \gamma^2(v, \Lambda^{-1}v).\end{aligned}$$

with $v = (c\times\Lambda^{-1}c)/(\|c\times\Lambda^{-1}c\|)$. Lemma 6.13 provides the justification to consider only nonnegative vectors c for which a little manipulation shows that

$$c\times\Lambda^{-1}c = \left(c_2c_3(\frac{1}{\lambda_3} - \frac{1}{\lambda_2}), c_1c_3(\frac{1}{\lambda_1} - \frac{1}{\lambda_3}), c_1c_2(\frac{1}{\lambda_2} - \frac{1}{\lambda_1})\right)^T \tag{6.44}$$

and

$$\operatorname{sgn}\,(\Lambda^{-2}c, v) = -\operatorname{sgn}\,((\lambda_2-\lambda_1)(\lambda_3-\lambda_1)(\lambda_3-\lambda_2)) = -1,$$

where we assume $c_1c_2c_3 \neq 0$. Otherwise, by (6.44) we would have $(\Lambda^{-2}c, v) = 0$ and $\bar{M}$ would not be modified by interchanging d_+ and d_-, since only $\bar{m}_{22}$ depends on v). Then the choice of these vectors would become meaningless.

Since $\bar{m}_{22}$ is positive for either choice of $d_\pm$ we sum up

$$0 < \bar{m}_{22}[d_+] < \bar{m}_{22}[d_-].$$

Now the spectral radius of $\bar{M}$ is a monotone increasing function in $\bar{m}_{22}$ because of

$$\frac{d}{d\bar{m}_{22}} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} \bar{m}_{11} & \bar{m}_{12} \\ \bar{m}_{21} & \bar{m}_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_2^2 \geq 0$$

for $(x_1, x_2)^T \in \mathbb{R}^2$. We have reached our goal since we identify d_+ as the search direction being responsible for the smaller spectral radius of $\bar{M}$ whose inverse leads to the larger one of each of the smaller Ritz values of $(\bar{A}, \bar{M})$ for $d_\pm$.

Let us now summarize the result in Theorem 6.20.

Theorem 6.20 (Poorest preconditioning). *Let $c \in \mathbb{R}^n$, $c \neq 0$, not collinear to any of the unit vectors. Lemma 6.13 provides the justification to assume $c \geq 0$. Assuming that Conjecture 6.16 holds, then PINVIT(2), as defined by Equation (6.40) for all admissible preconditioners, $\gamma \in [0, 1)$, is of poorest convergence (minimal decrease of the Rayleigh quotient concerning the choice of the preconditioner) within the search direction*

$$d_+ = (1-\gamma^2)^{1/2} \frac{\lambda\Lambda^{-1}c - c}{\|(I-\lambda\Lambda^{-1})c\|} + \gamma \frac{c \times \Lambda^{-1}c}{\|c \times \Lambda^{-1}c\|}. \tag{6.45}$$

The Rayleigh-Ritz procedure (6.33) and (6.34), by applying to the column space of $V = [c, d_+]$, provides an upper bound for the poorest decrease of the Rayleigh quotient.

In order to derive a convergence estimate for PINVIT(2) we are left with the task to determine the particular vector c^* of poorest convergence from

$$L^{[3]}(\lambda) = \{c \in \operatorname{span}\{e_1, e_2, e_n\} : \; \lambda(c) = \lambda, \; c \geq 0\}. \tag{6.46}$$

Despite a nonzero scaling constant there is a unique representation of $L^{[3]}(\lambda)$ by $c(\varphi)$, $\varphi \in (0, \pi/2)$ with the nonzero components

$$\begin{aligned} c_1 &= \left(\left(\frac{1}{\lambda} - \frac{1}{\lambda_2} \right) \cos^2(\varphi) + \left(\frac{1}{\lambda} - \frac{1}{\lambda_n} \right) \sin^2(\varphi) \right)^{1/2}, \\ c_2 &= \left(\frac{1}{\lambda_1} - \frac{1}{\lambda} \right)^{1/2} \cos(\varphi), \\ c_n &= \left(\frac{1}{\lambda_1} - \frac{1}{\lambda} \right)^{1/2} \sin(\varphi), \end{aligned} \tag{6.47}$$

for $\lambda \in [\lambda_1, \lambda_2]$, which can easily be checked by direct computation. Applying Theorem 6.20 to $c(\varphi)$ allows us to construct the case of poorest convergence (for all preconditioners obeying the spectral bound defined by γ). Finally, maximization in φ by a numerical method for given eigenvalues will provide an explicit upper bound for the worst-case convergence of PINVIT(2) on $L^{[3]}(\lambda)$.

Let us sum up the convergence properties in terms of the x-basis notation. Assume $x \in \mathbb{R}^n$ and let $\lambda_i < \lambda(x) < \lambda_{i+1}$ for $i < n-1$. Then in the case of exact preconditioning with $B^{-1} = A^{-1}$ it holds the asymptotically sharp bound by Theorem 6.3

$$\frac{\Delta_{i,i+1}(\lambda(x'))}{\Delta_{i,i+1}(\lambda(x))} \leq \left(\frac{\lambda_i(1 - \dfrac{\lambda_{i+1}}{\lambda_n})}{2\lambda_{i+1} - \lambda_i(1 + \dfrac{\lambda_{i+1}}{\lambda_n})} \right)^2 =: \sigma^2[INVIT(2)], \tag{6.48}$$

which turns into a sharp estimate as $\lambda \to \lambda_i$. For $\gamma \in (0, 1)$ we have a trivial upper bound by the PINVIT convergence theory

$$\frac{\Delta_{i,i+1}(\lambda(x'))}{\Delta_{i,i+1}(\lambda(x))} \leq \left(\gamma + (1-\gamma)\frac{\lambda_i}{\lambda_{i+1}} \right)^2 =: \sigma^2[PINVIT(1)], \tag{6.49}$$

while a sharp bound can be determined numerically by Theorem 6.20 if applied to $L^{[3]}(\lambda)$ by (6.46), which can be parametrized in $\varphi \in [0, \pi/2]$ according to (6.47). This section is closed with some conjecture on the upper bound for PINVIT(2). Numerical evidence will be given in Section 6.5.4.

Conjecture 6.21. *For $x \in \mathbb{R}^n$ and $\lambda_1 < \lambda(x) < \lambda_2$ the PINVIT(2)-iterate x' satisfies the (non-sharp) estimate*

$$\frac{\Delta_{1,2}(\lambda(x'))}{\Delta_{1,2}(\lambda(x))} \leq \left(\gamma + (1-\gamma^2)\sigma[INVIT(2)] \right)^2. \tag{6.50}$$

Obviously, (6.50) turns into a sharp estimate for $\gamma \to 0$ or $\gamma \to 1$.

6.5 Numerical algorithms

The aim of this section is twofold. On the one hand, the convergence factors of PINVIT(2) are illustrated in comparison with those of PINVIT(1). On the other hand, numerical evidence shall be given for the validity of Conjectures 6.16 and 6.21.

6.5.1 Numerical results for the Laplacian

We illustrate the convergence estimates derived above by computing these bounds for the 5 point finite difference approximation of the Laplacian on $[0, \pi]^2$ with homogeneous Dirichlet

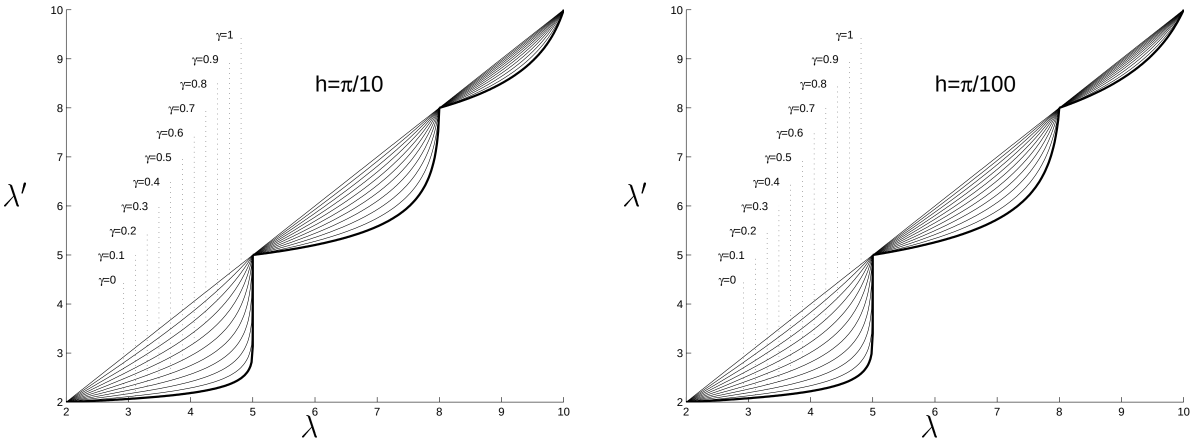


Figure 6.6: *Largest value of* λ' *for PINVIT(2) on* $L(\lambda)$ *with* $\lambda \in [1, 10]$. *Eigenvalues:* $(2, 5, 8, 10, \ldots, 4/h^2)$. *Left:* $h = \pi/10$. *Right:* $h = \pi/1000$.

γ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\lambda', h = \pi/10$	2.19	2.29	2.42	2.57	2.74	2.94	3.16	3.38	3.60	3.81	4
$\lambda', h = \pi/1000$	2.22	2.34	2.47	2.63	2.81	3.01	3.22	3.43	3.64	3.83	4

Table 6.1: λ' *for* $\lambda = 4$ *and* $\gamma = 0, 0.1, \ldots, 1.0$.

boundary conditions on the entire boundary. Using a uniform mesh with the size $h = \pi/N$, $N \in \mathbb{N}$, in both directions we obtain the discrete eigenvalues

$$\lambda^{(k,l)} = \frac{2}{h^2}\left(2 - \cos(kh) - \cos(lh)\right), \qquad k, l \in \mathbb{N},\ 1 \leq k, l \leq N - 1. \tag{6.51}$$

While the PINVIT convergence estimates on $L(\lambda)$ only depend on the nearest eigenvalues enclosing λ, we are in the case of PINVIT(2) faced with the necessity to define additionally the largest eigenvalue λ_n, i.e. $\lambda_n \approx 4/h^2$ for large N by (6.51). Equations (6.12) and (6.13) suggest, at least for a neighborhood of $\gamma = 0$, only a weak dependence on λ_n inasmuch h is sufficiently small. For the following examples we use $\lambda_n \approx 400/\pi^2$ for $h = \pi/10$ and $\lambda_n \approx 4 \cdot 10^6/\pi^2$ for $h = \pi/1000$. For simplicity the smallest eigenvalues are in both cases set to $(\lambda_1, \cdots, \lambda_4) = (2, 5, 8, 10)$, which are the limit values for $h \to 0$. In Figure 6.6 the largest Rayleigh quotient λ' attainable by PINVIT(2) is plotted against $\lambda \in [2, 10]$ for $\gamma = 0, 0.1, \ldots, 1$. In other words, the content of Figure 6.6 is an upper bound for the decrease of the Rayleigh quotient if PINVIT(2) is applied to an arbitrary vector in $L(\lambda)$. The reason why all curves intersect at $\lambda = \lambda_i$ is simply that $L(\lambda_i)$ contains the eigenvector associated with λ_i in which PINVIT(2) attains its poorest convergence, i.e. stationarity. The computations underlying this figure make use of Theorem 6.20 which is applied in the interval $[\lambda_i, \lambda_{i+1}]$ to the eigenvalues $(\lambda_i, \lambda_{i+1}, \lambda_n)$. Numerical maximization within the parameter φ, see Equation

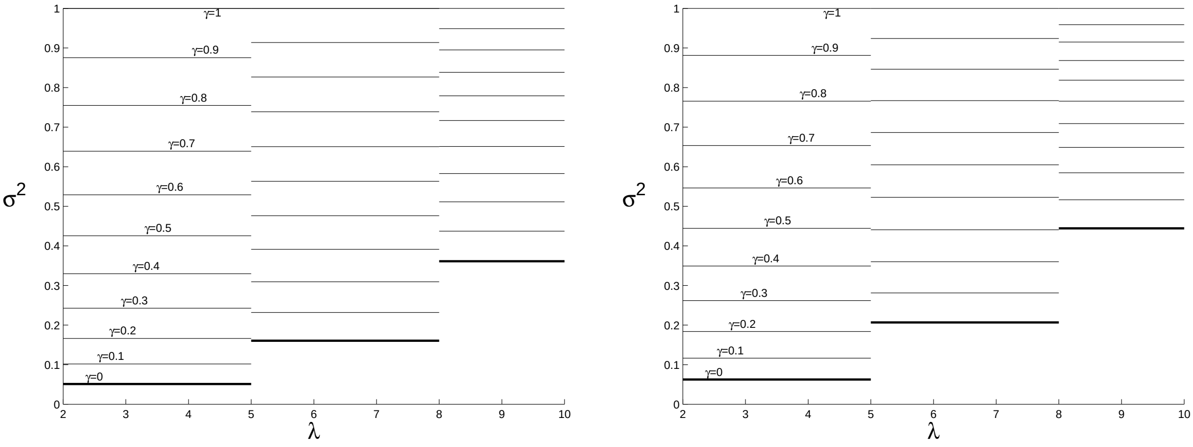


Figure 6.7: *Convergence factors σ^2 for PINVIT(2) on $L(\lambda)$ with $\lambda \in [1, 10]$. Eigenvalues: $(2, 5, 8, 10, \ldots, 4/h^2)$. Left: $h = \pi/10$. Right: $h = \pi/1000$.*

(6.47), serves to determine the vector of poorest convergence in $L(\lambda)$. The curves in Figure 6.6 for $h = \pi/10$ and $h = \pi/1000$ resemble each other, giving evidence for the weak dependence on λ_n; this is supported by the numerical data listed in Table 6.1.

The data presented in Figure 6.6 can also be used to compute an upper bound for

$$\Delta_{i,i+1}(\lambda')/\Delta_{i,i+1}(\lambda)$$

giving the convergence factors $\sigma^2[PINVIT(2)]$. These factors are displayed in Figure 6.7. Once more there is only a weak dependence on the choice of λ_n. Obviously, $\gamma = 1$ results in $\sigma[PINVIT(2)] = 1$ or stationarity of PINVIT(2). In the opposite case of exact preconditioning or $\gamma = 0$, Theorem 6.3 holds.

These data allow a comparison of the convergence factors of PINVIT(1) and PINVIT(2); see Figure 6.8. The PINVIT(1) convergence bounds $\sigma[PINVIT(1)]$ are drawn left as broken lines for $\gamma = 0, 0.1, \ldots, 1.0$. These broken lines have a constant value in $[2, 5]$ for each γ and they are by Theorem 4.1 asymptotically sharp upper estimates for the ratios

$$\Delta_{1,2}(\lambda')/\Delta_{1,2}(\lambda),$$

which are also plotted as dotted lines against $\lambda \in [\lambda_1, \lambda_2] = [2, 5]$.

There is a third type of curves shown in Figure 6.8: The solid curves stand for the worst case convergence of PINVIT(2). They are computed once more for $\gamma = 0, 0.1, \ldots, 1.0$ with an algorithm elucidated in Section 6.5.3, which is not built on Conjecture 6.16 so that we obtain further evidence for its validity by the numerical data.

Figure 6.8 (right side) also displays the convergence estimates

$$\Phi(\lambda, \gamma) = (\lambda' - \lambda_1)/(\lambda - \lambda_1)$$

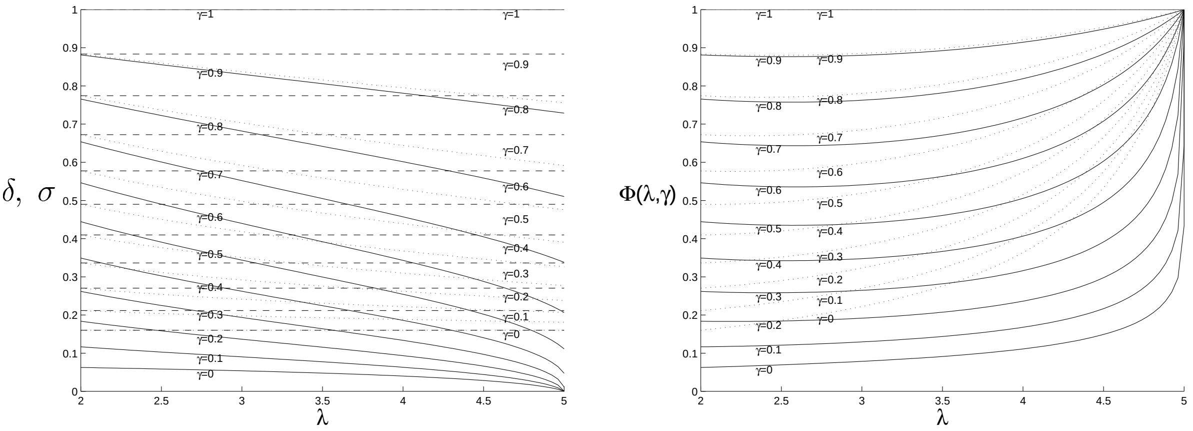


Figure 6.8: *Comparison of PINVIT and PINVIT(2) estimates in* $[\lambda_1, \lambda_2] = [2, 5]$ *with* $\lambda_n = 4/h^2$*,* $h = \pi/10^3$*. Left:* $\delta := \Delta_{1,2}(\lambda')/\Delta_{1,2}(\lambda)$*, Right:* $\Phi(\lambda, \gamma)$*.*

γ	0	0.1	0.2	0.3	0.4	0.5	0.7	0.9	1
$\sigma^2[PINVIT(1)]$	0.160	0.212	0.270	0.336	0.410	0.490	0.672	0.884	1.0
$\sigma^2[PINVIT(2)]$	0.062	0.117	0.184	0.262	0.349	0.444	0.654	0.881	1.0

Table 6.2: *Convergence factors* σ *for PINVIT and PINVIT(2).*

for PINVIT(2) by solid lines, and that for PINVIT(1) by using dotted lines. While the worst case convergence of PINVIT(1) is always bounded from below by the INVIT(1) convergence factor $\sigma^2[INVIT(1)] = (\lambda_1/\lambda_2)^2 = 0.16$ (the limit for $\gamma \to 0$), PINVIT(2) hastens convergence significantly.

For exact preconditioning, or $\gamma = 0$, the convergence factor $\sigma^2[INVIT(2)]$ by Equation (6.48) equals about 0.0625. This small value of this bounding factor $\sigma^2[INVIT(2)]$ provides a better justification for the convenient rule-of-thumb, namely that the convergence rate of a (multigrid) preconditioner transfers to that of PINVIT(2), since the “offset-factor” (i.e. 0.0625 for $\gamma = 0$) in the case of exact preconditioning is very small. Table 6.2 lists some numerical values for these σ^2 factors of PINVIT(1) and PINVIT(2).

Finally, note that the asymptotic convergence rates $\Phi(\lambda_1, \gamma)$ coincide with the aforementioned quantities since

$$\lim_{\lambda \to \lambda_1} \frac{\Delta_{1,2}(\lambda')}{\Delta_{1,2}(\lambda)} = \lim_{\lambda \to \lambda_1} \Phi(\lambda),$$

because of $\lim_{\lambda \to \lambda_1} (\lambda_2 - \lambda)/(\lambda_2 - \lambda') = 1$.

Let us summarize that acceleration techniques like PINVIT(2) have their largest impact on hastening poorest convergence for small γ (which should also be understood as a general principle for methods of the class of preconditioned subspace iterations). In contrast to this, for γ near 1, PINVIT(1) as well as PINVIT(2) reach stationarity in the most unfavorable case.

But in practice, PINVIT(1) as well as PINVIT(2) open a wide corridor between poorest and best convergence, cf. Chapter 3. In fact, for an appropriate choice of the preconditioner the fastest convergence is achieved for $\gamma = 1$, since $E_1(c)$ is the largest set of new PINVIT iterates and $E_1(c) - c$ defines the largest set of search directions in the case of PINVIT(2).

6.5.2 Connectedness of $L_+(\lambda)$

The next lemma shows that the level set of nonnegative vectors with a fixed Rayleigh quotient is arcwise-connected and therefore underpins the applicability of the numerical search algorithms on $L_+(\lambda)$ presented in Section 6.5.3. We note that the level set $L(\lambda)$ (containing all $c \in \mathbb{R}^n$ with $\lambda(c) = \lambda$) is a non-connected set, as no path of a constant Rayleigh quotient through the origin can be constructed.

Lemma 6.22. *Let $\lambda \in (\lambda_1, \lambda_n)$. The level set of the Rayleigh quotient on all nonnegative vectors*

$$L_+(\lambda) = \{c \in \mathbb{R}^n \,:\, c \geq 0,\ \lambda(c) = \lambda\}$$

is arcwise-connected.

Proof. The proof is only given for the case of simple eigenvalues but its generalization is straightforward. We first note that $L_+(\lambda)$ is radially connected, i.e. $\kappa c \in L_+(\lambda)$ for any $\kappa > 0$. Therefore, we take no notice of the norm of the paths we construct. Let us start with $\lambda \neq \lambda_i$ and assume $\lambda \in (\lambda_i, \lambda_{i+1})$. We define $\bar{c} \in L_+(\lambda)$ with only two nonzero components given by

$$\bar{c}_i = \left(\frac{\lambda_i(\lambda_{i+1} - \lambda)}{\lambda(\lambda_{i+1} - \lambda_i)}\right)^{1/2}, \qquad \bar{c}_{i+1} = \left(\frac{\lambda_{i+1}(\lambda - \lambda_i)}{\lambda(\lambda_{i+1} - \lambda_i)}\right)^{1/2}. \tag{6.52}$$

We show now that any $c \in L_+(\lambda)$ is joined with $\bar{c}$ by an arc in $L_+(\lambda)$.

If $c \neq \bar{c}$, then there is a positive component c_j for $j \neq i, i+1$. First assume $j < i$. In order to find a path from c to $\bar{c}$ we damp out the component c_j (by multiplying c_j^2 with $0 \leq \epsilon \leq 1$) and compensate this by increasing c_i^2 (by adding $\eta > 0$). This process takes place on a path in $L_+(\lambda)$ since for any $\epsilon \in [0, 1]$ the equation

$$\lambda = \frac{\sum_{k\neq i,j}^n c_k^2 + \epsilon c_j^2 + (c_i + \eta)^2}{\sum_{k\neq i,j}^n c_k^2/\lambda_k + \epsilon c_j^2/\lambda_j + (c_i + \eta)^2/\lambda_i}$$

implies

$$(\epsilon - 1)(\frac{\lambda}{\lambda_j} - 1)c_j^2 = (\eta^2 + 2\eta c_i)(\frac{\lambda}{\lambda_i} - 1)$$

and has the positive solution $\eta = \eta(\epsilon)$

$$\eta = \left(c_i^2 + (1 - \epsilon)c_j^2 \frac{\lambda\lambda_j^{-1} - 1}{\lambda\lambda_i^{-1} - 1}\right)^{1/2} - c_i.$$

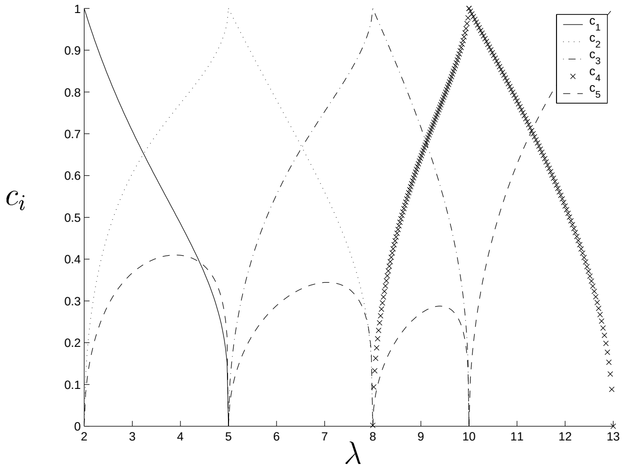


Figure 6.9: *Coefficients of the vector of poorest PINVIT(2) convergence*

The endpoint $\epsilon = 0$ of the arc is a vector with a vanishing jth component. We proceed similarly if $j > i+1$, but now we increase the ith component. Having removed all components different from i and $i+1$ we end in a multiple of $\bar{c}$.

The remaining case $\lambda = \lambda_i$ is treated as follows: To show that any $c \in L_+(\lambda_i)$ is connected in $L_+(\lambda_i)$ with $\bar{c}$, we first ensure $c_i > 0$ by continuously increasing the ith component, which does not change the Rayleigh quotient. Then we follow the path $\epsilon \in [0, 1]$

$$c(\epsilon, \eta) := (\epsilon c_1, \ldots, \epsilon c_{i-1}, c_i, \eta c_{i+1}, \ldots, \eta c_n) \in L_+(\lambda_i),$$

where $\eta = \eta(\epsilon)$ is given by

$$\eta = \epsilon \left((\sum_{k=1}^{i-1} c_k^2 (\frac{\lambda_i}{\lambda_k} - 1)) / (\sum_{k=i+1}^{n} c_k^2 (1 - \frac{\lambda_i}{\lambda_k})) \right)^{1/2}.$$

□

6.5.3 A search algorithm on $L_+(\lambda)$

In order to give numerical evidence for the validity of Conjecture 6.16, we are looking for an algorithm to determine numerically the vector of poorest convergence of PINVIT(2) on $L(\lambda)$ for given $\gamma \neq 0$. Hence, we have to deal with two nested constrained optimization problems where the objective functional λ' is the Rayleigh quotient of the PINVIT(2) iterate, which is maximized.

1. Outer Loop: Search routine on $L_+(\lambda)$. We perform a sequential random search with line minimization. The underlying idea is essentially the same as the one used in the proof of

Max. dev.	d_1	d_2	d_3	d_4	d_5
$[2, 5]$	$3.55 \cdot 10^{-5}$	$7.34 \cdot 10^{-5}$	$9.77 \cdot 10^{-4}$	$8.66 \cdot 10^{-4}$	$1.01 \cdot 10^{-4}$
$[5, 8]$	$2.71 \cdot 10^{-5}$	$6.03 \cdot 10^{-5}$	$3.53 \cdot 10^{-4}$	$2.87 \cdot 10^{-3}$	$3.93 \cdot 10^{-4}$
$[8, 10]$	$1.89 \cdot 10^{-5}$	$6.66 \cdot 10^{-5}$	$3.60 \cdot 10^{-4}$	$7.89 \cdot 10^{-4}$	$9.89 \cdot 10^{-4}$

Figure 6.10: *Deviation between “experimental” and “theoretical” coefficients.*

Lemma 6.22. To begin with, for given $c \in L_+(\lambda)$ with $\lambda \in (\lambda_i, \lambda_{i+1})$, two nonnegative, nonzero random vectors

$$a = (a_1, \ldots, a_i, 0, \ldots, 0)^T, \qquad b = (0, \ldots, 0, b_{i+1}, \ldots, b_n)^T$$

are generated. Then $\lambda(c + a) < \lambda$ and $\lambda(c + b) > \lambda$. By the mean value theorem for given $\alpha > 0$ a certain $\beta = \beta(\alpha) > 0$ always exists with

$$\lambda(c + \alpha a + \beta b) = \lambda.$$

A line search in α is performed to maximize the objective functional λ', i.e. the Rayleigh quotient of the PINVIT(2) iterate of $\tilde{c} := c + \alpha a + \beta(\alpha) b$.

2. Inner loop: Preconditioned steepest descent is applied to $\tilde{c}$. Obviously, Conjecture 6.16 is not applied. But owing to Theorem 6.18, cf. Remark 6.19, and instead of presenting the search space as $V = [\tilde{c}, (I + UDU^T)(\tilde{c} - \lambda(\tilde{c})\Lambda^{-1}\tilde{c})]$ for *any* orthogonal $U \in \mathbb{R}^{n\times n}$ and diagonal D, $-\gamma \leq d_{ii} \leq \gamma$, one can confine to considering

$$V = [\tilde{c}, (I + \gamma u u^T)(\tilde{c} - \lambda(\tilde{c})\Lambda^{-1}\tilde{c})]$$

for all $u \in \mathbb{R}^n$ on the unit ball.

The program code to solve the nested optimization problem is written in MATLAB. For the inner loop the MATLAB template `sg_min` of Edelman et. al. [37, 38] has been applied which realizes the optimization with respect to the Stiefel manifold $Stief(n, k)$ of $n \times k$ orthogonal matrices (here $k = 1$). We only provide the PINVIT(2) functional, and a dog-leg step algorithm (interpolating steepest descent and a Newton's method step) was selected where the Euclidean metric endows the constraint surface.

For a numerical illustration of Conjecture 6.16 we have selected (the low-dimensional) example matrix $\Lambda = \mathrm{diag}(2, 5, 8, 10, 13)$. Our approach to check the validity of Conjecture 6.16 consists of two steps:

(A) In the first step we generate reference data on the basis of Conjecture 6.16. This is done by applying PINVIT(2) to the vector $c(\varphi)$ as defined by Equation (6.47) where φ parametrizes $L^{[3]}(\lambda)$. PINVIT(2) in the interval $[\lambda_1, \lambda_4] = [2, 10]$ is applied to these

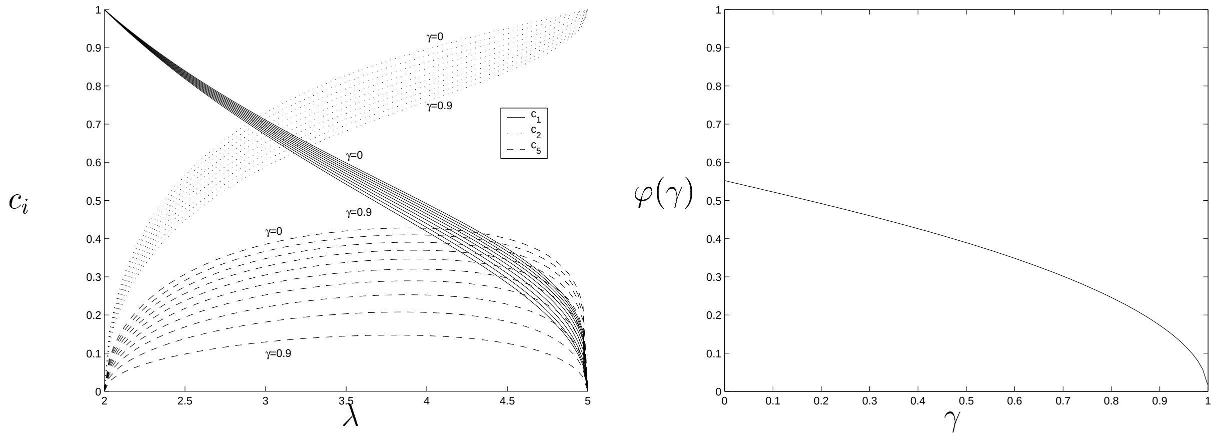


Figure 6.11: *γ-dependence of vector of poorest PINVIT(2) convergence. Left:*c_1*,* c_2*,* c_5 *in* $[\lambda_1, \lambda_2] = [2, 5]$ *, Right:* $\varphi^*(\gamma)$*.*

$c(\varphi)$ and Theorem 6.20 is used to determine the poorest convergence in dependence on the choice of the preconditioner. In a second outer loop, the angle $\varphi^* \in (0, \pi/2)$ for given λ is determined so that PINVIT(2) attains its poorest convergence on $L(\lambda)$. The components of $c(\varphi^*)$, embedded in the $\mathbb{R}^5$, are drawn in Figure 6.9.

Finally, in the interval $[\lambda_4, \lambda_5] = [10, 13]$ poorest convergence of PINVIT(2) is fully controlled by Lemma 6.12. The components of the vector of poorest convergence are given by (6.52) for $i = 4$.

(B) In a second step the search algorithm on $L_+(\lambda)$, as elucidated at the beginning of this section, is applied. Therefore, the interval $[\lambda_1, \lambda_{4]}$ is subdivided into 400 equidistant grid points and on each level set $L(\lambda)$ the vector of poorest convergence is determined *without* using Conjecture 6.16.

The numerical results gained in steps (A) and (B) strikingly confirm Conjecture 6.16. After a number of 100 sweeps of the outer search loop on $L_+(\lambda)$ the deviation between the "theoretical" coefficients (see (A)) and those "experimental" coefficients (see (B)) is small. Table 6.10 lists the deviations as defined by

$$d_i := \max_{\lambda \in [\lambda_j, \lambda_{j+1}]} |c_i^{\text{Theory}} - c_i^{\text{Exp.}}|, \qquad i = 1, \ldots, 5, \quad j = 1, 2, 3.$$

Finally, we would like to point out that the components of $c(\varphi^*)$ by (6.47) depend on γ. This is to be seen in contrast to the behavior of PINVIT(1) (cf. [95] and Chapter 3) where the corresponding vector is γ-independent. For our test problem Figure 6.11 shows the components c_1, c_2 and c_5 in $[\lambda_1, \lambda_2]$ for various γ; additionally the $\varphi^*[\gamma]$ is drawn.

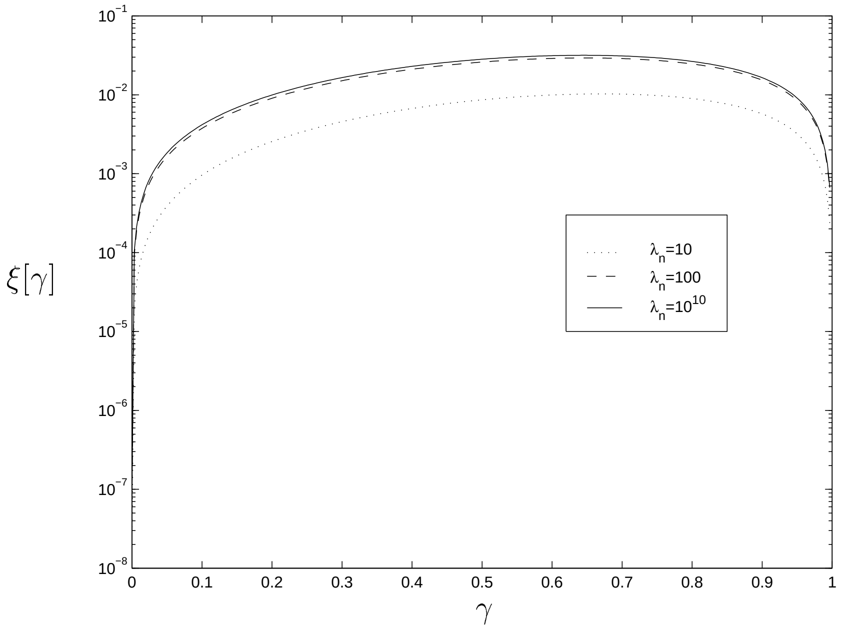


Figure 6.12: $\xi[\gamma]$ *for* $\lambda_n = 10,\ 100,\ 10^{10}$.

6.5.4 On conjecture 6.21

We check Conjecture 6.21 numerically within the $\mathbb{R}^3$ for $\Lambda = \operatorname{diag}(2, 5, \lambda_3)$, and $\lambda_3 = 10$, 100, 10^{10}. For each Λ the quantity

$$\xi[\gamma] := \hat{\sigma}^2[PINVIT(2)] - \left(\max_{\lambda\in[2,5]} \frac{\Delta_{i,i+1}(\lambda')}{\Delta_{i,i+1}(\lambda)} \right)$$

with

$$\hat{\sigma}[PINVIT(2)] := \gamma + (1-\gamma^2)\sigma[INVIT(2)]$$

is computed and plotted in Figure 6.12 versus $\gamma \in [0, 1]$. Since $\xi[\gamma]$ is positive, Conjecture 6.21 is never violated. Obviously, $\sigma^2[PINVIT(2)]$ is not a sharp estimate, but according to the theory, it is asymptotically sharp as $\gamma \to 0$ or $\gamma \to 1$.

6.6 Critical conclusion

- A new convergence theory of INVIT(2) and central elements of such a theory for PINVIT(2) have been presented.

- The new estimate for INVIT(2) is sharp in λ_i, λ_{i+1}, γ and asymptotically sharp in λ.

- The geometry of PINVIT(2) has been cleared up as well as the dependence of poorest convergence on the choice of the preconditioner.
- Poorest convergence with respect to the level set $L(\lambda)$ is currently not underpinned by theoretical results. Conjecture 6.16 makes a mini-dimensional convergence analysis of PINVIT(2) possible.
- The numerical tests give every evidence for the validity of Conjecture 6.16.

Geometric methods as used for the analysis of PINVIT(k), $k = 1, 2$, seem to be very successful for understanding the underlying principles of these preconditioned eigensolvers. It is an open question how comparable techniques can help to understand the most important and very promising scheme of PINVIT(3), also called LOPCG, see Section 7.1.

7. Numerical experiments

In this chapter we report on the results of some numerical experiments with the PINVIT schemes. Firstly, in Section 7.1, our intention is to give a numerical comparison of the efficiency of PINVIT(k,s) for small k and various preconditioners. This will provide numerical evidence for the optimality of PINVIT(3,s), also called LOBPCG [72]. These demonstrations are restricted to our model problem, i.e. the Laplacian on $[0, \pi]^2$, discretized by using linear finite elements. Numerical comparison of PINVIT(k,s) with the Rayleigh quotient multigrid minimization technique of Mandel and McCormick [84] and the direct multigrid approach of Hackbusch [52, 55] are contained in [93], while a comparison with the Jacobi-Davidson type schemes JDQR and JDCG is given in [74].

Subsequently, in Section 7.2 we present an adaptive scheme for PINVIT(k,s). Adaptive discretization methods are well known to provide numerical solutions of partial differential equations and corresponding eigenproblems (within some prescribed tolerance) with only a small portion of the work necessary when uniform grid refinement is employed. The necessary iteration error estimator and discretization error estimator are briefly reviewed [97]. Numerical experiments on a slit domain with mixed boundary conditions exemplify the effectiveness of these a posteriori error estimators.

7.1 Comparison of the PINVIT(k,s) schemes

We consider the eigenproblem for the Laplacian on $[0, \pi]^2$ with homogeneous Dirichlet boundary conditions. The problem is discretized by using linear finite elements on a uniform triangle mesh with the grid parameter $h = \pi/64$ and 3969 inner nodes. (Numerical results for a larger problem with more than 16×10^6 nodes are contained in [93].) The discretized eigenproblem is a generalized matrix eigenvalue problem for the pencil $A - \lambda M$, where A is the stiffness matrix for the Laplacian and M is the mass matrix. The PINVIT(k,s) schemes are tested for the following multigrid preconditioners: on the one hand, we apply a V-cycle preconditioner [14] using Gauss-Seidel smoothing (alternatively Jacobi smoothing) on a hierarchy of grids $h_l = \pi/2^l$, $l = 2, \ldots, 6$, and exact solution on the coarsest grid. On the other hand, we make use of the hierarchical basis preconditioner [144, 147] acting on grids with h_l, $l = 1, \ldots, 6$, so that the coarsest finite element space only consists of a single basis function.

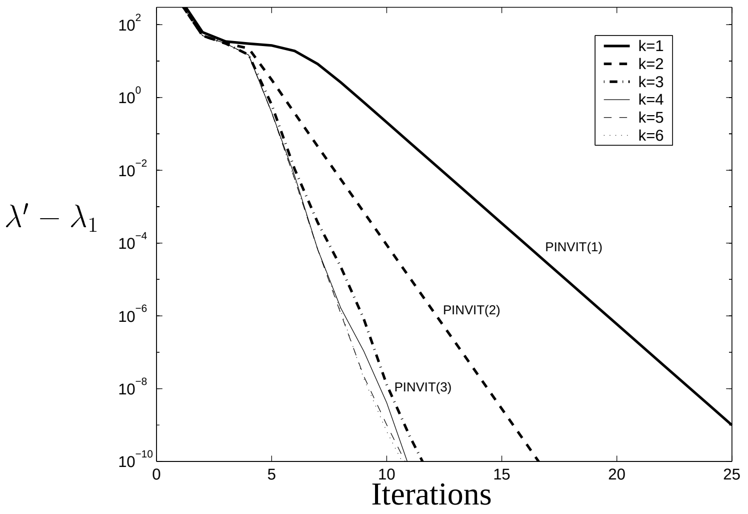


Figure 7.1: *Convergence of* $\lambda' - \lambda_1$ *of PINVIT(k) for* $k = 1, \ldots, 6$.

Experiment 1. To begin with, we compare the results of PINVIT(k,1), $k = 1, \ldots, 6$ where we use the $V(2,2)$-cycle preconditioner with 2 steps of Jacobi pre- and postsmoothing each. Figure 7.1 displays the error $\lambda' - \lambda_1$ of the computed eigenvalue approximations λ' versus the iteration index for PINVIT(k,s) and $k = 1, \ldots, 6$. In fact, each curve represents the case of poorest convergence toward λ_1 for 100 random initial vectors (the same for each k). The relatively poor convergence in the first steps accounts for $\lambda' > \lambda_2$ and the attraction to eigenvalues larger than λ_1. The outcome of this experiment exemplifies (for $s = 1$) that

1. PINVIT(1,s) and PINVIT(2,s) are both less efficient than PINVIT(3,s) (which even holds if related to the total expense per step),

2. PINVIT(3,s) or LOBPCG appears as the optimal scheme since the slope of the curves for PINVIT(k,s), $k \geq 4$, is more or less the same as the one of PINVIT(3,s).

The optimality of PINVIT(3,s) was first described by Knyazev, see [70, 72, 73], but a theoretical analysis is still not available. In [72] Knyazev points out that PINVIT(3,s) has a *conjugate gradient like convergence behavior* and calls the scheme Locally Orthogonal Block Preconditioned Conjugate Gradient (LOBPCG). The numerical experiments in [72, 75] strikingly confirm the cg-like convergence properties.

In the light of this optimality of PINVIT(3,s), we restrict the further discussion to the schemes PINVIT(k,s), $k = 1, 2, 3$.

Experiment 2. Next, we compare the computed convergence factors for the vector schemes PINVIT(k), $k = 1, 2, 3$, with the theoretical factors for PINVIT(1), which are trivial upper

PINVIT(k)	$V(1,1)$				$V(2,2)$			
	$\bar{\sigma}^2$	$\sigma^2_{\max}$	γ_{est}	σ^2_{Theory}	$\bar{\sigma}^2$	$\sigma^2_{\max}$	γ_{est}	σ^2_{Theory}
$k=1$	0.167	0.202	0.399	0.408	0.155	0.170	0.119	0.221
$k=2$	0.122	0.254	0.306	0.340	0.063	0.0876	0.065	0.193
$k=3$	0.106	0.215	0.317	0.348	0.025	0.062	0.074	0.197

Table 7.1: Convergence factors for PINVIT(k,1) for $k = 1, 2, 3$.

bounds for PINVIT(k), $k \geq 2$. This is done for $V(i,i)$-cycle preconditioning performing $i = 1, 2$ steps of Gauss-Seidel pre- and postsmoothing. The results are listed in Table 7.1.

Each scheme is applied to the same 200 random initial vectors and the iteration is stopped if the error $\lambda' - \lambda_1$ is less than 10^{-8}. All schemes for each random initial vector converge to the smallest eigenpair.

The quantities listed in Table 7.1 are defined as follows: $\bar{\sigma}^2$ is the mean value of all convergence factors σ^2 (see Chapter 4) computed from the numerical data by

$$\frac{\lambda' - \lambda_1}{\lambda_2 - \lambda'} \cdot \frac{\lambda_2 - \lambda}{\lambda - \lambda_1} =: \sigma^2.$$

The σ-factors are only recorded if $\lambda < \lambda_2$ and $\sigma_{\max}$ is the maximum over all these σ for the iterations to all 200 initial vectors. The quantity γ_{est} is the maximum of all ratios of residuals (taken in the Euclidean norm)

$$\frac{\|Ax' - \lambda Mx\|}{\|Ax - \lambda Mx\|},$$

where, once more, the ratios are only stored if $\lambda < \lambda_2$. Hence, γ_{est} defines the maximal eigenvalue of the error propagation matrix $I - B^{-1}A$ "seen" by the iterates. Finally, σ^2_{Theory} is computed by inserting γ_{est} in (4.17), the convergence factor of PINVIT(1).

Figure 7.2 displays the convergence history of PINVIT(k) for the $V(2,2)$ preconditioner. Therein, $\lambda' - \lambda_1$ and σ^2 are plotted versus the iteration index for 20 random initial vectors. The slope of the bold line in Figure 7.2(a) represents the theoretical asymptotic behavior of PINVIT(1) which is determined by σ_{Theory}, i.e., we have drawn $(\sigma_{\text{Theory}})^{2i}(\lambda_2 - \lambda_1)$ against the iteration index i. In contrast to this, the bold line in Figure 7.2(b) is the theoretical bound as derived in Theorem 4.6 for PINVIT(1), i.e. $(\sigma(\gamma_{\text{est}}, \lambda_1, \lambda_2))^2$ with $\gamma_{\text{est}} \approx 0.119$ corresponding to PINVIT(1) with $V(2,2)$ preconditioning. Note that the first iteration is of a relatively fast convergence and that the convergence later slows down.

Table 7.2 lists the analogous data for the hierarchical basis preconditioner. The iteration is terminated whenever $\lambda' - \lambda_1$ is less than 10^{-6}. Since the hierarchical basis preconditioner only satisfies (2.4) for some $\delta_1 > 2$, we have enforced PINVIT(1) convergence by scaling the preconditioner with the constant $1/2$. For this choice we observe the poor value $\gamma_{\text{est}} \approx 0.94$.

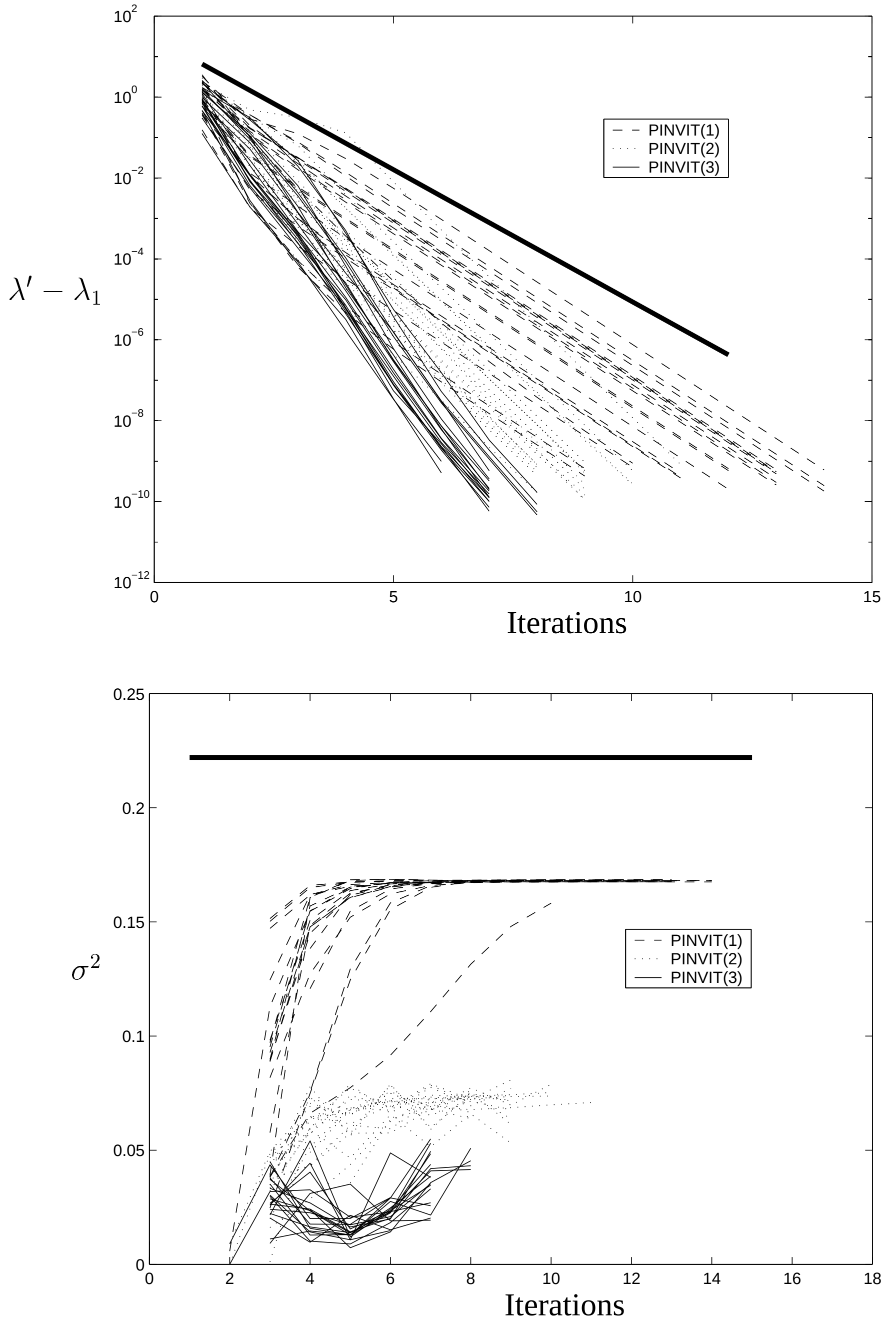


Figure 7.2: $\boxed{\substack{a\\b}}$ *PINVIT(k) convergence for $V(2,2)$-preconditioning recorded for 20 initial vectors each. (a) $\lambda' - \lambda_1$ against the iteration index. (b) The convergence factor σ^2.*

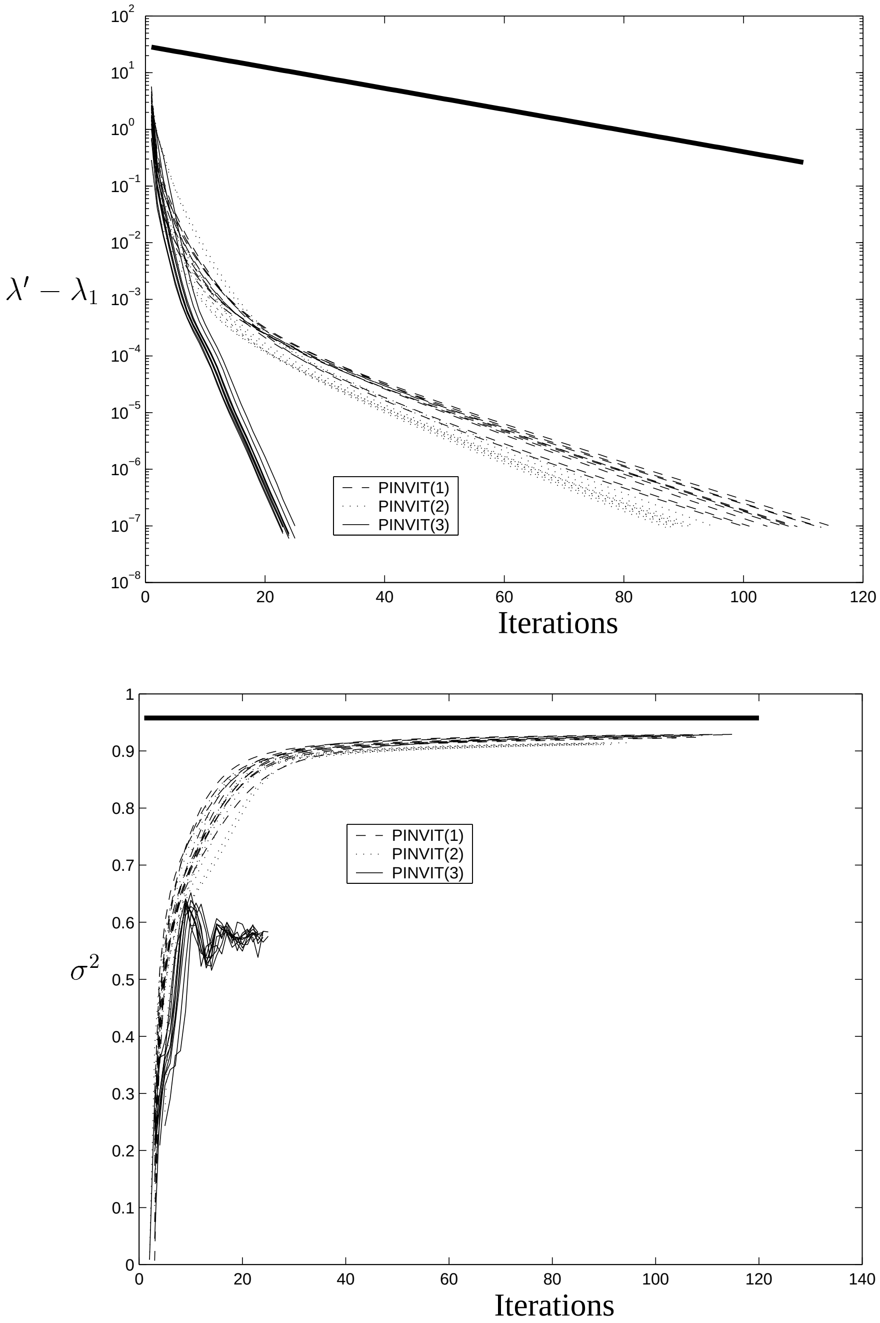


Figure 7.3: $\begin{array}{|c|}\hline a \\ b \\ \hline\end{array}$ *PINVIT(k), $k = 1, 2, 3$, convergence for hierarchical basis preconditioning recorded for 10 initial vectors each. (a) $\lambda' - \lambda_1$ against the iteration index. (b) The convergence factor σ^2.*

	$\bar{\sigma}^2$	$\sigma^2_{\max}$
PINVIT(1)	0.883	0.929
PINVIT(2)	0.868	0.916
PINVIT(3)	0.519	0.633

Table 7.2: PINVIT(k) using hierarchical basis preconditioning.

		Residuals		
i	j	PINVIT(1,s))	PINVIT(2,s))	PINVIT(3,s))
1	1	1.91×10^{-6}	1.91×10^{-6}	1.91×10^{-6}
1	2	4.17×10^{-1}	4.17×10^{-1}	4.17×10^{-1}
1	3	4.14×10^{-1}	4.14×10^{-1}	4.14×10^{-1}
1	4	4.32×10^{-1}	4.32×10^{-1}	4.32×10^{-1}
3	1	1.24×10^{-3}	9.38×10^{-4}	8.31×10^{-4}
3	2	7.65×10^{-3}	9.22×10^{-3}	5.51×10^{-3}
3	3	3.31×10^{-2}	2.25×10^{-2}	2.10×10^{-2}
3	4	3.49×10^{-2}	4.84×10^{-2}	3.66×10^{-2}
6	1	2.32×10^{-6}	3.86×10^{-7}	2.95×10^{-7}
6	2	4.32×10^{-4}	1.22×10^{-5}	6.23×10^{-6}
6	3	9.43×10^{-4}	2.20×10^{-4}	6.75×10^{-5}
6	4	5.94×10^{-3}	1.22×10^{-2}	7.74×10^{-3}

Table 7.3: *PINVIT(k,s): Residuals in step $i = 1, 3, 6$ for the jth Ritz vector.*

The corresponding convergence history is presented in Figure 7.4 providing evidence for the excellent performance of PINVIT(3).

Experiment 3. We consider the preconditioned subspace scheme PINVIT(k,s) for $V(2,2)$ preconditioning with Gauss-Seidel smoothing. The 7-dimensional initial subspace is constructed in a way that the jth column of V is given as the grid restriction of the function $(x/\pi)^{j/2} + (y/\pi)^{j/3}$. Rayleigh-Ritz is applied to V and results in the initial $V^{(0)}$. After this, PINVIT(k,s) is applied to $V^{(0)}$ for $k = 1, 2, 3$. In Figure 7.4 the differences $\lambda_j^{(i)} - \lambda_j$ are recorded for $j = 1, \ldots, 4$, versus the iteration index i. The iteration is stopped if $\lambda_4^{(i)} - \lambda_4 \leq 10^{-8}$. This is the case after 23 PINVIT(1,7) iterations but only 10 PINVIT(3,7) steps. In order to allow easy comparison all plots have been scaled to the same abscissa. Finally, Table 7.3 lists the residuals (with respect to the Euclidean norm) of the M-normalized Ritz vectors for the first iterations.

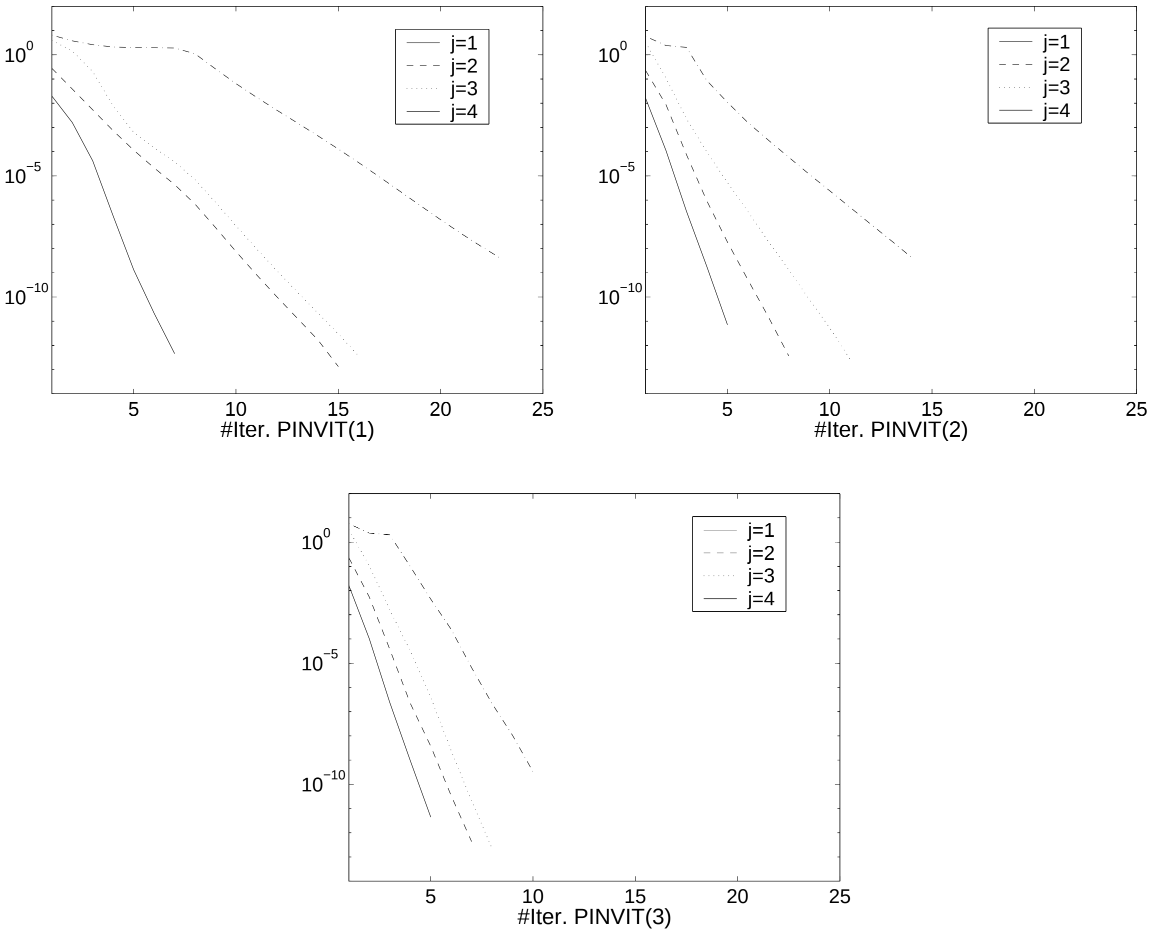


Figure 7.4: *Preconditioned subspace iteration:* $\lambda_j^{(i)} - \lambda_j$ *for* $j = 1, \ldots, 4$ *versus the iteration index* i *for PINVIT(k,7) using* $V(2,2)$ *preconditioning.*

7.2 An adaptive subspace eigensolver

Let us report in condensed form on an adaptive scheme for PINVIT(k,s). Central elements of an adaptive solver, both for boundary value problems as well as for eigenvalue problems of a given partial differential operator, are *a posteriori* error estimators. A posteriori error estimators for the eigenvalue problem can be derived in the general setup of Galerkin methods for nonlinear variational problems [131]. Those estimators, specifically designed for the eigenvalue problem, have not reached the same attention as the ones for boundary value problems; see the works of Friberg [43], Heuveline and Rannacher [59] as well as [97].

In [97] a residual-based a posteriori error estimator has been suggested. By using the notation of Section 5.1 the estimator F_i reads

$$F_i = (\nabla\lambda(v_i), B^{-1}d_i) = 2\|d_i\|_B^2, \tag{7.1}$$

where v_i is the ith column of V, i.e. the ith Ritz vector, and d_i is the associated preconditioned residual.

As shown in Theorem 3.1 in [97], F_i provides an upper estimate for the distance of the ith Ritz value θ_i, $\theta_i \in [\lambda_m, \lambda_{m+1})$, to the nearest eigenvalues λ_m and λ_{m+1} in the form

$$(\theta_i - \lambda_m)(\lambda_{m+1} - \theta_i) \leq \frac{\lambda_{m+1}}{2(1-\gamma)} F_i. \tag{7.2}$$

The Courant-Fischer principles guarantee a saturation assumption (cf. [10]) to hold for F_i.

Equation (7.2) can be used to define an *iteration error estimator*, in order to construct a stopping condition for the iterative solver, as well as a *discretization error estimator* providing local indicators to steer the mesh refinement process. The latter is a hierarchical estimator, since the error indicators are evaluated within a space of higher order elements. The error estimator for PINVIT(1,s), as suggested in [97], can also be applied to any of the schemes PINVIT(k,s). Both error estimators can be coupled to a certain subset $P \subseteq \{1, \ldots, s\}$ containing indexes of eigenfunctions of low regularity. The adaptive multigrid eigensolver should give best results for those eigenfunctions whose indexes are contained in P.

We reproduce from [92, 97] the numerical study of the eigenproblem for the Laplacian on a slit annulus; implementational details are described in [80, 81]. Therefore, let the domain be given by

$$\Omega_r = \{z \in \mathbb{R}^2 \,:\, r \leq \|z\| \leq 1\} \setminus A^+,$$

where A^+ denotes the axis $(x, 0)^T$, $x > 0$. Homogeneous Dirichlet boundary conditions are supposed on the circles with the radii r_0 and 1, while homogeneous Dirichlet (Neumann) boundary conditions are given on the top (bottom) of the slit. In Figure 7.5 we show for $r_0 = 0$ the contour lines of the first three eigenfunctions together with some final triangulations for various choices of P. Analogous results for $r_0 = 1/4$ are depicted in Figure 7.6. The analytical solutions of both problems in terms of Bessel functions of the first and second kind and fractional order are given in [92, 97], where one can also find further information on the accuracy of the numerical solutions, which illustrate the efficacy of the adaptive multigrid scheme.

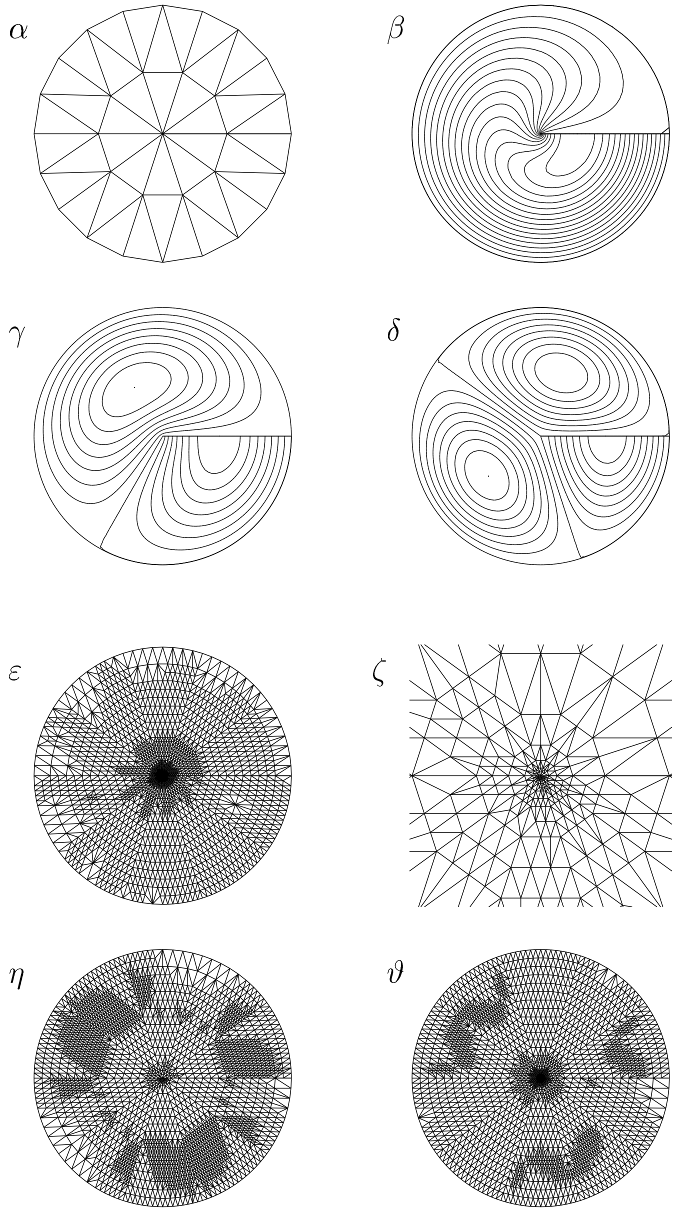


Figure 7.5: *(α) Initial triangulation. (β, γ, δ) Contour plots of* $u_{0,1}$, $u_{1,1}$, $u_{2,1}$. *Final triangulations: (ε, ζ) $P = \{1\}$, 2385 nodes and zoom of* $[-2^{-10}, 2^{-10}]^2$. *(η) $P = \{3\}$, 2374 nodes. (ϑ) $P = \{1, 2, 3\}$, 2381 nodes.*

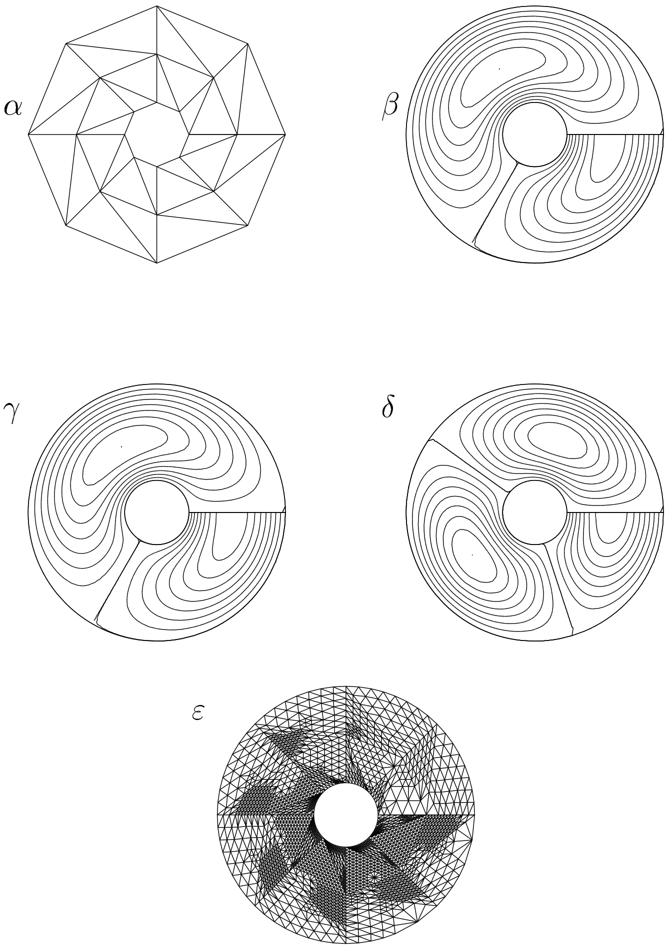


Figure 7.6: *(α) Initial triangulation. (β, γ, δ) Contour plots of* $u_{0,1}$, $u_{2,1}$, $u_{3,1}$. *(ε) Final triangulation, $P = \{1\}$, 2141 nodes.*

8. Conclusion and outlook

In this work we have presented an innovative approach to the analysis and classification of preconditioned gradient type eigensolvers. These eigensolvers (mainly developed for the solution of symmetric positive definite eigenvalue problems deriving from the finite difference or finite element discretization of self-adjoint and coercive elliptic partial differential operators) have been known in the literature since the late 1950s. The "classical" convergence analysis of these preconditioned eigensolvers developed in the last decades has only led to *asymptotically* sharp convergence estimates.

The present work gives a reinterpretation of preconditioned gradient type eigensolvers and of some of their extensions and generalizations within the theoretical setup of (a variant of) subspace iteration, modified in such a way that the associated linear system is solved approximately by using preconditioning. On this basis, a certain *hierarchy of preconditioned eigensolvers* is suggested, in which the classical preconditioned gradient scheme appears as the most simple representative. Within the light of our reinterpretation we prefer to call this scheme preconditioned inverse iteration.

New proof techniques have been devised within this framework, leading to *non-asymptotic sharp* estimates for preconditioned inverse iteration. Such estimates have not only been derived concerning the *poorest* convergence, but additionally those estimates have been provided, which describe the *fastest possible* convergence. On the one hand, the estimates on the poorest convergence constitute a considerable improvement of the classical estimates, see [95]. On the other hand, the estimates on the fastest possible convergence promote an understanding for the extremely fast convergence of such schemes, which is often observed in the first iteration steps. Moreover, several sharp estimates have been derived for the iterative schemes of steepest descent and preconditioned steepest descent, which are called INVIT(2) and PINVIT(2) within our hierarchy of preconditioned eigensolvers.

The analytic treatment of the preconditioned eigensolvers discussed in this work was originally stimulated by the *underlying geometry*. The key point is that for the basic scheme of preconditioned inverse iteration the set of possible iterates (corresponding to all admissible preconditioners) is given by a ball with respect to the A-geometry. This intrinsic geometry has been proved very useful for carrying out the analysis. However, the symmetry and positive definiteness of A are decisive for our setup and seem to limit the applicability of these proof techniques to larger classes of eigenproblems, e.g. for non-symmetric matrices.

The geometric interpretation has also conveyed a deepened insight into preconditioning for iterative eigensolvers. The resulting and increased understanding of preconditioning techniques for the iterative solution of the eigenproblem proves these eigensolvers as

+ *stable and robust*: Convergence to an eigenpair from scratch is guaranteed even for poor preconditioners.

+ *conceptionally simple*: Only some matrix-vector multiplications are to be provided as "black-box" routines, namely the products with the discretization and mass matrix and with the preconditioner. There is no necessity to hold any of these operators as full matrices in the computer storage. Beyond that, no matrix factorizations are to be performed.

  In contrast to that, only simple linear operations, inner products and the Rayleigh-Ritz procedure are required to realize the PINVIT(k,s) schemes.

+ *easy to implement and cheap*: Any multigrid solver as developed for the solution of boundary value problems can be interpreted as a preconditioner and can be employed within a preconditioned eigensolver. Therefore, no elaborate programming techniques are required in order to write special multigrid code for the eigenproblem. Each iteration step can be realized with $\mathcal{O}(n)$ operations in the best case. Not too much effort should be made to construct very accurate preconditioners, since overly accurate approximate inverses do not always guarantee fastest convergence to an eigenpair.

As a very important feature of preconditioned eigensolvers, *grid-independent convergence* holds for such high-quality multigrid preconditioners which are equipped with mesh independent estimates on their quality. In addition to this, by using the advanced *multi-level preconditioners*, multigrid convergence for eigensolvers can be shown *for the first time* under the assumptions made in the theory of multilevel preconditioning: *Preconditioned eigensolvers can solve eigenproblems on non-uniform grids and without any assumptions on the regularity of the problem with optimal computational complexity.*

We hope that the geometric techniques introduced in this work will prove as useful tools for the analysis of the following (partially) unsolved problems as given within the setup of our mesh eigenproblems:

- a convergence analysis for the Locally Optimal Preconditioned Conjugate Gradient method [70, 72], which we call PINVIT(3,s) within the hierarchy of preconditioned eigensolvers introduced in Chapter 1. Such a theory should also prove the optimality and cg-like convergence properties of PINVIT(3,s) compared to all schemes PINVIT(k,s) for $k \geq 4$.

- a sound theory of preconditioned eigensolvers using shift strategies in order to compute eigenpair approximations from the interior of the spectrum without previous knowledge of any smaller eigenvalues. Such an approach will have to deal with the difficulty of preconditioning indefinite matrices.

- a (geometric) understanding of those improved preconditioning strategies, which allow us (in the sense of Chapter 3) to construct highly efficient preconditioners especially for the eigenvalue problem. Those preconditioners will be poor for the solution of linear systems, but should increase the speed of convergence toward the searched eigenpair.

- the further development and convergence theory of those refined preconditioned eigensolvers, which are built on a defect correction scheme in the orthogonal complement of the actual eigenvector approximation.

- the construction of efficient preconditioned iterative solvers for quadratic eigenproblems, see [5, Chapter 9] and [128], which, e.g., occur in the vibration analysis of damped mechanical structures.

REFERENCES

[1] G.P. Astrakhantsev. A certain iterative method of solution of network elliptic problems. *USSR Comput. Math. and Math. Physics*, 11:171–182, 1971.

[2] G.P. Astrakhantsev. The iterative improvement of eigenvalues. *USSR Comput. Math. and Math. Physics*, 16,1:123–132, 1976.

[3] I. Babuška and J. Osborn. *Handbook of numerical Analysis*, volume II, chapter Eigenvalue problems. Elsevier, North–Holland, 1991.

[4] N.S. Bahvalov. Convergence of a relaxation method under natural constraints on an elliptic operator. *USSR J. Comput. Math. and Math. Physics*, 6:861–883, 1966.

[5] 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*. SIAM, Philadelphia, 2000.

[6] R. E. Bank and T. Dupont. An optimal order process for solving finite element equations. *Math. Comp.*, 36(153):35–51, 1981.

[7] R.E. Bank. Analysis of a multilevel inverse iteration procedure for eigenvalue problems. *SIAM J. Numer. Anal.*, 19:886–898, 1982.

[8] F.L. Bauer. Das Verfahren der Treppeniteration und verwandte Verfahren zur Lösung algebraischer Eigenwertprobleme. *ZAMP*, 8:214–235, 1957.

[9] T.L. Beck. Real-space mesh techniques in density functional theory. *Rev. Mod. Phys.*, 72:1041–1080, 2000.

[10] F.A. Bornemann, B. Erdmann, and R. Kornhuber. A posteriori error estimation for elliptic problems in two and three space dimensions. *SIAM J. Numer. Anal.*, 33:1188–1204, 1996.

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

[12] D. Braess. The contraction number of a multigrid method for solving the Poisson equation. *Numer. Math.*, 37(3):387–404, 1981.

[13] D. Braess and W. Hackbusch. A new convergence proof for the multigrid method including the V-cycle. *SIAM J. Numer. Anal.*, 20(5):967–975, 1983.

[14] J.H. Bramble. *Multigrid methods*. Pitman Research Notes in Mathematics Series. Longman, London, 1993.

[15] 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.

[16] J.H. Bramble, J.E. Pasciak, J.P. Wang, and J. Xu. Convergence estimates for multigrid algorithms without regularity assumptions. *Math. Comp.*, 57(195):23–45, 1991.

[17] J.H. Bramble, J.E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comp.*, 55(191):1–22, 1990.

[18] A. Brandt. Multi-level adaptive solutions to boundary-value problems. *Math. Comp.*, 31(138):333–390, 1977.

[19] A. Brandt. Multiscale computation in chemistry. Technical report, Department of computer science and applied mathematics, The Weizmann institute of science, Rehovot, Israel, 1999.

[20] A. Brandt, S. McCormick, and J. Ruge. Multigrid methods for differential eigenproblems. *SIAM J. Sci. Statist. Comput.*, 4(2):244–260, 1983.

[21] V.E. Bulgakov, M.V. Belyi, and K.M. Mathisen. Multilevel aggregation method for solving large-scale generalized eigenvalue problems in structural dynamics. *Internat. J. Numer. Methods Engrg.*, 40(3):453–471, 1997.

[22] Z. Cai, J. Mandel, and S. McCormick. Multigrid methods for nearly singular linear equations and eigenvalue problems. *SIAM J. Numer. Anal.*, 34:178–200, 1997.

[23] T.F. Chan and I. Sharapov. Subspace correction multilevel methods for elliptic eigenvalue problems. CAM Report 96–34, Dept. of Math., Univ. of Calif., Los Angeles, 1996.

[24] F. Chatelin. *Eigenvalues of matrices*. Wiley, Chichester, 1993.

[25] L. Collatz. *Eigenwertprobleme und ihre numerische Behandlung*. Akademische Verlagsgesellschaft, Leipzig, 1945.

[26] R. Courant and D. Hilbert. *Methoden der mathematischen Physik*. Springer, Berlin, 1924.

[27] E.R. Davidson. The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices. *J. Comput. Phys.*, 17:87–94, 1975.

[28] P. Deuflhard, T. Friese, and F. Schmidt et. al. Effiziente Eigenmodenberechnung für den Entwurf integriert-optischer Chips. Scientific report 96–2, Konrad-Zuse-Zentrum für Informationstechnik Berlin, 1996.

[29] P.A.M. Dirac. Quantum mechanics of many-electron systems. *Proc. Roy. Soc. London A*, 123:714–733, 1929.

[30] B. Döhler. Ein neues Gradientenverfahren zur simultanen Berechnung der kleinsten oder größten Eigenwerte des allgemeinen Eigenwertproblems. *Numer. Math.*, 40(1):79–91, 1982.

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

[32] E.G. D'yakonov. *Optimization in solving elliptic problems*. CRC Press, Boca Raton, Florida, 1996.

[33] E.G. D'yakonov and A. V. Knyazev. Group iterative method for finding lower-order eigenvalues. *Moscow Univ. Comput. Math. Cybern.*, 2:32–40, 1982.

[34] E.G. D'yakonov and A. V. Knyazev. On an iterative method for finding lower eigenvalues. *Russian J. Numer. Anal. Math. Modelling*, 7(6):473–486, 1992.

[35] E.G. D'yakonov and M.Y. Orekhov. Minimization of the computational work in eigenvalue problems. *Dokl. Akad. Nauk SSSR*, 235(5):1005–1008, 1977.

[36] 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.

[37] A. Edelman, T.A. Arias, and S.T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20:303–353, 1999.

[38] A. Edelman and R. Lippert. Nonlinear eigenvalue problems with orthogonality constraints. 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*. SIAM, Philadelphia, 2000.

[39] D. K. Faddeev and V. N. Faddeeva. *Computational methods of linear algebra*. W. H. Freeman and Co., San Francisco, 1963.

[40] J.L. Fattebert and J. Bernholc. Towards grid-based O(N) density-functional theory methods: Optimized nonorthogonal orbitals and multigrid acceleration. *Phys. Rev. B*, 62(3):1713–1722, 2000.

[41] R.P. Fedorenko. A relaxation method for the solution of elliptic partial differential operators. *USSR J. Comput. Math. and Math. Physics*, 1,5:1092–1096, 1961. (In Russian).

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

[43] P.O. Friberg. An error indicator for the generalized eigenvalue problem using the hierarchical finite element method. *Int. J. Numer. Meth. Engrg.*, 23:91–98, 1986.

[44] T. Friese, P. Deuflhard, and F. Schmidt. A multigrid method for the complex Helmholtz eigenvalue problem. In *Eleventh International Conference on Domain Decomposition Methods (London, 1998)*, pages 18–26 (electronic). DDM.org, Augsburg, 1999.

[45] F.R. Gantmacher. *Matrizentheorie*. Springer, Berlin, 1986.

[46] 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.

[47] G. H. Golub and H. A. van der Vorst. Eigenvalue computation in the 20th century. *J. Comput. Appl. Math.*, 123(1-2):35–65, 2000. Numerical analysis 2000, Vol. III. Linear algebra.

[48] G.H. Golub and C.F. Van Loan. *Matrix computations*. John Hopkins University Press, Baltimore, MD, 3rd edition, 1996.

[49] G.H. Golub and Q. Ye. Inexact inverse iteration for generalized eigenvalue problems. *BIT*, 40(4):671–684, 2000.

[50] H. van der Vorst. Subspace iteration for eigenproblems. *CWI Quarterly*, 9(1-2):151–160, 1996.

[51] W. Hackbusch. Ein iteratives Verfahren zur schnellen Auflösung elliptischer Randwertprobleme. Report 76-12, Mathematisches Institut der Universität zu Köln, 1976.

[52] W. Hackbusch. On the computation of approximate eigenvalues and eigenfunctions of elliptic operators by means of a multi-grid method. *SIAM J. Numer. Anal.*, 16:201–215, 1979.

[53] W. Hackbusch. Convergence of multigrid iterations applied to difference equations. *Math. Comp.*, 34:425–440, 1980.

[54] W. Hackbusch. On the convergence of multigrid iterations. *Beiträge zur numerischen Mathematik*, 9:213–239, 1981.

[55] W. Hackbusch. *Multi-grid methods and applications*. Springer series in computational mathematics 4. Springer, Berlin, 1985.

[56] W. Hackbusch and G. Hofmann. Results of the eigenvalue problem for the plate equation. *Z. Angew. Math. Phys.*, 31(6):730–739, 1980.

[57] T. Helgaker, P. Jørgensen, and J. Olsen. *Molecular electronic-structure theory*. John Wiley & Sons, Chichester, 2000.

[58] 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.

[59] V. Heuveline and R. Rannacher. A posteriori error control for finite element approximations of elliptic eigenvalue problems. SFB 359 Preprint 8/2001, Universität Heidelberg, 2001.

[60] R. Hiptmair and K. Neymeyr. Multilevel method for mixed eigenproblems. Sonderforschungsbereich 382, Universitäten Tübingen und Stuttgart, Report 159, Submitted to SIAM J. Sci. Comp., 2001.

[61] T. Hwang and I. D. Parsons. A multigrid method for the generalized symmetric eigenvalue problem. I. Algorithm and implementation. *Internat. J. Numer. Methods Engrg.*, 35(8):1663–1676, 1992.

[62] T. Hwang and I. D. Parsons. A multigrid method for the generalized symmetric eigenvalue problem. II. Performance evaluation. *Internat. J. Numer. Methods Engrg.*, 35(8):1677–1696, 1992.

[63] 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.

[64] C.G.J. Jacobi. Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen. *Journal für die reine und angewandte Mathematik; Crelles Journal*, 30:51–94, 1846.

[65] L. V. Kantorovich. *Functional analysis and applied mathematics*. U. S. Department of Commerce National Bureau of Standards, Los Angeles, Calif., 1952. Translated by C. D. Benster.

[66] L.V. Kantorovich and G.P. Akilov. *Functional analysis in normed spaces*. Pergamon, London, 1964.

[67] A.V. Knyazev. Private communication. 2000.

[68] A.V. Knyazev. Computation of eigenvalues and eigenvectors for mesh problems: algorithms and error estimates. (In Russian), Dept. Numerical Math., USSR Academy of Sciences, Moscow, 1986.

[69] A.V. Knyazev. Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem. *Russian J. Numer. Anal. Math. Modelling*, 2:371–396, 1987.

[70] A.V. Knyazev. A preconditioned conjugate gradient method for eigenvalue problems and its implementation in a subspace. In *International Ser. Numerical Mathematics, 96, Eigenwertaufgaben in Natur- und Ingenieurwissenschaften und ihre numerische Behandlung, Oberwolfach, 1990.*, pages 143–154, Basel, 1991. Birkhäuser.

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

[72] A.V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM J. Sci. Comp.*, 23:517–541, 2001.

[73] A.V. Knyazev and K. Neymeyr. A geometric theory for preconditioned inverse iteration. III: A short and sharp convergence estimate for generalized eigenvalue problems. *Accepted for Linear Algebra Appl.*, 2001.

[74] A.V. Knyazev and K. Neymeyr. A geometric theory for preconditioned inverse iteration. III: A short and sharp convergence estimate for generalized eigenvalue problems. Technical Report 161, Sonderforschungsbereich 382, Universitäten Tübingen und Stuttgart,, 2001.

[75] A.V. Knyazev and K. Neymeyr. Efficient solution of symmetric eigenvalue problems using multigrid preconditioners in the locally optimal block conjugate gradient method. Technical Report 163, Sonderforschungsbereich 382, Universitäten Tübingen und Stuttgart, Accepted for ETNA, 2001.

[76] 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.

[77] Y. Lai, K. Lin, and W. Lin. An inexact inverse iteration for large sparse eigenvalue problems. *Numer. Linear Algebra Appl.*, 4:425–437, 1997.

[78] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Stand.*, 45:255–282, 1950.

[79] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK users' guide*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods.

[80] P. Leinen. Data structures and concepts for adaptive finite element meshes. *Computing*, 55:325–354, 1995.

[81] P. Leinen, W. Lembach, and K. Neymeyr. An adaptive subspace method for elliptic eigenproblems with hierarchical basis preconditioning. Technical report, Sonderforschungsbereich 382, Universitäten Tübingen und Stuttgart, 1997.

[82] W. Lembach. *Adaptive Mehrgitterverfahren zur Lösung von Eigenwertproblemen elliptischer Differentialoperatoren*. PhD thesis, Universität Tübingen, 1998.

[83] D.E. Longsine and S.F. McCormick. Simultaneous Rayleigh-quotient minimization methods for $Ax = \lambda Bx$. *Linear Algebra Appl.*, 34:195–234, 1980.

[84] J. Mandel and S.F. McCormick. A multilevel variational method for $Au = \lambda Bu$ on composite grids. *J. Comput. Phys.*, 80:442–452, 1989.

[85] S.F. McCormick. A general approach to one-step iterative methods with application to eigenvalue problems. *J. Comp. Sys. Sci.*, 6:354–372, 1972.

[86] S.F. McCormick. Some convergence results on the method of gradients for $Ax = \lambda Bx$. *J. Comp. Sys. Sci.*, 13:213–222, 1976.

[87] S.F. McCormick. A mesh refinement method for $Ax = \lambda Bx$. *Math. Comp.*, 36(154):485–498, 1981.

[88] S.F. McCormick. Multilevel adaptive methods for elliptic eigenproblems: a two-level convergence theory. *SIAM J. Numer. Anal.*, 31:1731–1745, 1994.

[89] A. Meyer. *Modern algorithms for large sparse eigenvalue problems*, volume 34 of *Mathematical Research*. Akademie-Verlag, Berlin, 1987.

[90] R. B. Morgan and D. S. Scott. Preconditioning the Lanczos algorithm for of sparse symmetric eigenvalue problems. *SIAM J. Sci. Comput.*, 14(3):585–593, 1993.

[91] R.B. Morgan. Davidson's method and preconditioning for generalized eigenvalue problems. *J. Comp. Phys.*, 89:241–245, 1990.

[92] K. Neymeyr. A posteriori error estimation for a preconditioned algorithm to solve elliptic eigenproblems. Technical Report 77, Sonderforschungsbereich 382, Universitäten Tübingen und Stuttgart, 1997.

[93] K. Neymeyr. Solving mesh eigenproblems with multigrid efficiency. Technical Report 157, Sonderforschungsbereich 382, Universitäten Tübingen und Stuttgart, 2000.

[94] K. Neymeyr. Why preconditioning gradient type eigensolvers? Sonderforschungsbereich 382, Universitäten Tübingen und Stuttgart, Report 146, (revised version), 2000.

[95] K. Neymeyr. A geometric theory for preconditioned inverse iteration. I: Extrema of the Rayleigh quotient. *Linear Algebra Appl.*, 322:61–85, 2001.

[96] K. Neymeyr. A geometric theory for preconditioned inverse iteration. II: Convergence estimates. *Linear Algebra Appl.*, 322:87–104, 2001.

[97] K. Neymeyr. A posteriori error estimation for elliptic eigenproblems. Sonderforschungsbereich 382, Universitäten Tübingen und Stuttgart, Report 132, (revised version), Accepted for J. Numer. Linear Algebra Appl., 2001.

[98] K. Neymeyr. A geometric theory for preconditioned inverse iteration applied to a subspace. *Math. Comp.*, 71:197–216, 2002.

[99] M.G. Neytcheva and P.S. Vassilevski. Preconditioning of indefinite and almost singular finite element elliptic equations. *SIAM J. Sci. Comput.*, 19(5):1471–1485, 1998.

[100] Y. Notay. Convergence analysis of inexact Rayleigh quotient iterations. Technical report, Service de Métrologie Nucléaire, Université Libre de Bruxelles, 2001.

[101] Y. Notay. Combination of Jacobi-Davidson and conjugate gradients for the partial symmetric eigenproblem. *Numer. Lin. Alg. Appl.*, 9:21–44, 2002.

[102] C. Ochsenfeld. Linear scaling exchange gradients for Hartree-Fock and hybrid density functional theory. *Chem. Phys. Lett.*, 327:216–223, 2000.

[103] S. Oliveira. On the convergence rate of a preconditioned subspace eigensolver. *Computing*, 63(3):219–231, 1999.

[104] E.E. Ovtchinnikov. Convergence estimates for the generalized Davidson method for the symmetric generalized eigenvalue problem. Technical report, Centre for Techno-Mathematics & Scientific Computing, London, 2001.

[105] E.E. Ovtchinnikov and L.S. Xanthis. Successive eigenvalue relaxation I: Theory. Technical Report Report 1-6-00, Centre for Techno-Mathematics & Scientific Computing, London, 2000.

[106] E.E. Ovtchinnikov and L.S. Xanthis. Successive eigenvalue relaxation: A new method for generalized eigenvalue problems and convergence estimates. *Proc. R. Soc. Lond. A*, 457:441–451, 2001.

[107] B.N. Parlett. *The symmetric eigenvalue problem*. Prentice Hall, Englewood Cliffs New Jersey, 1980.

[108] R.G. Parr and W. Yang. *Density-functional theory of atoms and molecules*. Oxford University Press, Oxford, 1989.

[109] G. Peters and J. H. Wilkinson. Inverse iteration, ill-conditioned equations and Newton's method. *SIAM Rev.*, 21(3):339–360, 1979.

[110] 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.

[111] V.G. Prikazchikov. Strict estimates of the rate of convergence of an iterative method of computing eigenvalues. *USSR J. Comput. Math. and Math. Physics*, 15:235–239, 1975.

[112] P.A. Raviart and J.-M. Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles*. Masson, Paris, 1992.

[113] G. Rodrigue. A gradient method for the matrix eigenvalue problem $Ax = \lambda Bx$. *Numer. Math.*, 22:1–16, 1973.

[114] A. Ruhe. Iterative eigenvalue algorithms based on convergent splittings. *J. Computational Phys.*, 19(1):110–120, 1975.

[115] A. Ruhe and T. Wiberg. The method of conjugate gradients used in inverse iteration. *Nordisk Tidskr. Informationsbehandling (BIT)*, 12:543–554, 1972.

[116] H.R. Rutishauser. Computational aspects of Bauer's simultaneous iteration method. *Numer. Math.*, 13:4–13, 1969.

[117] Y. Saad. *Numerical methods for large eigenvalue problems*. Manchester University Press, Manchester, 1992.

[118] Y. Saad. *Iterative methods for sparse linear systems*. PWS Publishing Company, Boston, 1996.

[119] 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).

[120] G. L. G. Sleijpen, A. G. L. Booten, D. R. Fokkema, and H. A. Van der Vorst. Jacobi-Davidson type methods for generalized eigenproblems and polynomial eigenproblems. *BIT*, 36(3):595–633, 1996. International Linear Algebra Year (Toulouse, 1995).

[121] 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.

[122] G.L.G. Sleijpen and F.W. Wubs. Effective preconditioning techniques for eigenvalue problems. Technical Report 1117, Universiteit Utrecht, Department of Mathematics, 1999.

[123] P. Smit and M. Paardekooper. The effects of inexact solvers in algorithms for symmetric eigenvalue problems. *Linear Algebra Appl.*, 287:337–357, 1999.

[124] J. Stoer and R. Bulirsch. *Numerische Mathematik 2*. Springer-Verlag, Berlin, 1990.

[125] L. G. Strakhovskaya and R. P. Fedorenko. On the solution of the principal spectral problem in the mathematical modeling of nuclear reactors. *Zh. Vychisl. Mat. Mat. Fiz.*, 40(6):920–928, 2000.

[126] L.G. Strakhovskaya. An iterative method for evaluating the first eigenvalue of an elliptic operator. *USSR Comput. Math. and Math. Physics*, 17,3:88–101, 1977.

[127] A. Szabo and N.S. Ostlund. *Modern quantum chemistry*. McGraw-Hill, New York, 1982.

[128] F. Tisseur and K. Meerbergen. The quadratic eigenvalue problem. *SIAM Rev.*, 43(2):235–286, 2001.

[129] H.A. van der Vorst and G.H. Golub. 150 years old and still alive: eigenproblems. In *The state of the art in numerical analysis (York, 1996)*, pages 93–119. Oxford Univ. Press, New York, 1997.

[130] P.S. Vassilevski. Preconditioning nonsymmetric and indefinite finite element matrices. *J. Numer. Linear Algebra Appl.*, 1:59–76, 1992.

[131] R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. Wiley and Teubner, New York and Stuttgart, 1995.

[132] E.L. Wachspress. *Iterative solution of elliptic systems, and applications to the neutron diffusion equations of reactor physics*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1966.

[133] J. Wang and T.L. Beck. Efficient real-space solution of the Kohn–Sham equations with multiscale techniques. *J. Chem. Phys.*, 112(21):9223–9228, 2000.

[134] S.R. White and J.W. Wilkins. Finite-element method for electronic structure. *Phys. Rev. B*, 39:5819–5833, 1989.

[135] H. Wielandt. Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme, Teil I: Abzählung der Eigenwerte komplexer Matrizen. Technical Report B43/J/9, Aerodynamische Versuchsanstalt Göttingen, Germany, 1943.

[136] H. Wielandt. Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme, Teil III: Das Iterationsverfahren in der Flatterrechnung. Technical Report B44/J/21, Aerodynamische Versuchsanstalt Göttingen, Germany, 1944.

[137] H. Wielandt. Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme, Teil IV: Konvergenzbeweis für das Iterationsverfahren. Technical Report B44/J/38, Aerodynamische Versuchsanstalt Göttingen, Germany, 1944.

[138] H. Wielandt. Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme, Teil V: Bestimmung höherer Eigenwerte durch gebrochene Iteration. Technical Report B44/J/38, Aerodynamische Versuchsanstalt Göttingen, Germany, 1944.

[139] H. Wielandt. Das Iterationsverfahren bei nicht selbstadjungierten linearen Eigenwertaufgaben. *Mathematische Zeitschrift*, 50:93–143, 1944. First appeared as Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme, Teil II, Bericht B43/J/21, Aerodynamische Versuchsanstalt Göttingen, Germany 1943.

[140] J. H. Wilkinson. Eigenvalue problems. In *The state of the art in numerical analysis (Birmingham, 1986)*, pages 1–39. Oxford Univ. Press, New York, 1987.

[141] J.H. Wilkinson, editor. *The algebraic eigenvalue problem*. Clarendon Press, Oxford, 1965.

[142] J.H. Wilkinson and C. Reinsch, editors. *Handbook for automatic computation*, volume II Linear Algebra. Springer, New York, 1971.

[143] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Rev.*, 34:581–613, 1992.

[144] H. Yserentant. On the multi-level splitting of finite element spaces. *Numer. Math.*, 49:379–412, 1986.

[145] H. Yserentant. On the multilevel splitting of finite element spaces for indefinite elliptic boundary value problems. *SIAM J. Numer. Anal.*, 23(3):581–595, 1986.

[146] H. Yserentant. Preconditioning indefinite discretization matrices. *Numer. Math.*, 54(6):719–734, 1989.

[147] H. Yserentant. Hierarchical bases. In *ICIAM 91 (Washington, DC, 1991)*, pages 256–276. SIAM, Philadelphia, PA, 1992.

[148] H. Yserentant. Old and new convergence proofs for multigrid methods. In *Acta numerica*, pages 285–326. Cambridge University Press, Cambridge, 1993.

[149] T. Zhang, G.H. Golub, and K.H. Law. Subspace iterative methods for eigenvalue problems. *Linear Algebra Appl.*, 294:239–258, 1999.

[150] P. F. Zhuk and L. N. Bondarenko. Sharp estimates for the rate of convergence of the s-step method of steepest descent in eigenvalue problems. *Ukraïn. Mat. Zh.*, 49(12):1694–1699, 1997.

List of Symbols

INVIT(k,s) 22
PINVIT(k,s) 23

$\mathcal{B}_\gamma$ 40
$C_\gamma(c)$ 55
$E_\gamma(x)$ 42
$E_\gamma^k(x)$ 42
$E_\gamma(c)$ 45
$F_\gamma(c)$ 116
$L(\lambda)$ 65
$L_+(\lambda)$ 127
$Q_\gamma(c)$ 117
$S_\gamma(c)$ 116
$Z_\gamma(c)$ 120

$\Delta_{p,q}(\lambda)$ 34
$\Pi(\cdot)$ 117

$c_{i,j}$ 65
$d_\pm$ 121
x_i 6

λ 8
λ_i 6
$\underline{\lambda}(c,\varphi)$ 69
$\lambda_{i,j}^\pm(\lambda,\gamma)$ 70
$\varphi_\gamma(c)$ 66
$\psi_\gamma(c)$ 66