Foundations of Computational Mathematics
https://doi.org/10.1007/s10208-025-09694-x

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Accuracy Controlled Schemes for the Eigenvalue Problem of the Radiative Transfer Equation

Wolfgang Dahmen[1] · **Olga Mula**[2]

Received: 21 August 2023 / Revised: 18 November 2024 / Accepted: 26 November 2024

Abstract

The criticality problem in nuclear engineering asks for the principal eigenpair of a Boltzmann operator describing neutron transport in a reactor core. Being able to reliably design, and control such reactors requires assessing these quantities within quantifiable accuracy tolerances. In this paper, we propose a paradigm that deviates from the common practice of approximately solving the corresponding spectral problem with a fixed, presumably sufficiently fine discretization. Instead, the present approach is based on first contriving iterative schemes, formulated in function space, that are shown to converge at a quantitative rate without assuming any a priori excess regularity properties, and that exploit only properties of the optical parameters in the underlying radiative transfer model. We develop the analytical and numerical tools for approximately realizing each iteration step within judiciously chosen accuracy tolerances, verified by a posteriori estimates, so as to still warrant quantifiable convergence to the exact eigenpair. This is carried out in full first for a Newton scheme. Since this is only locally convergent we analyze in addition the convergence of a power iteration in function space to produce sufficiently accurate initial guesses. Here we have to deal with intrinsic difficulties posed by compact but unsymmetric operators preventing standard arguments used in the finite dimensional case. Our main point is that we can avoid any condition on an initial guess to be already in a small neighborhood of the exact solution. We close with a discussion of remaining intrinsic obstructions to a certifiable numerical implementation, mainly related to not knowing the gap between the principal eigenvalue and the next smaller one in modulus.

Communicated by Endre Süli.

✉ Olga Mula
o.mula@tue.nl

[1] Department of Mathematics, University of South Carolina, Columbia, USA

[2] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

Published online: 21 January 2025

Keywords Eigenvalue problem · Spectral problem · Neutron transport · Radiative transfer · Error-controlled computation · A posteriori estimation · Iteration in function space

Mathematics Subject Classification 68Q25 · 68R10 · 68U05

1 Introduction

1.1 Context, Problem Formulation, and Main Objectives

Radiative transfer plays a key role in a number of scientific and engineering areas. It is, for example, relevant in understanding certain atmospheric processes and is also an essential constituent in modeling neutron transport in the context of nuclear engineering. The mathematical problem studied in this paper arises from this latter field. When operating a nuclear reactor, engineers seek to create a sustainable chain reaction in which the neutrons that are produced balance the neutrons that are either absorbed or leave the boundary. When the balance is achieved, the neutron population finds itself in a steady state which does not evolve in time. The specific nature of the balance and the neutron distribution crucially depend on the material and geometry of the reactor, hence the importance of studying this problem at design stages in order to find appropriate core configurations. In the field of nuclear engineering, this task is known as the criticality problem, and it is modelled through the computation of a generalized unsymmetric eigenvalue problem of a linear Boltzmann operator. Depending on the community, the process is also referred to as radiative transfer or neutron transport.

1.1.1 A Model Problem

To be specific, we consider the following model problem that nevertheless exhibits all essential obstructions: For a given bounded spatial domain $\mathrm{D} \subset \mathbb{R}^d$ (the nuclear reactor), we consider the population density of neutrons $u = u(x, v)$ located at position $x \in \mathrm{D}$ and travelling with velocity $v \in \mathrm{V} \subset \mathbb{R}^d$. The balance between the neutrons that are produced and the ones that are lost (by leakage or absorption) is mathematically expressed as the following generalized unsymmetric eigenvalue problem (usually called the criticality problem): the pair (u, λ) with $u \neq 0$, and $\lambda \in \mathbf{C}$ is said to be an eigenpair of the radiative transfer equation if and only if

$$\begin{aligned} v \cdot \nabla u(x, v) + \sigma(x, v)u(x, v) - \int_{\mathrm{V}} \kappa(x, v', v)u(x, v')\,\mathrm{d}v' \\ = \lambda \int_{\mathrm{V}} \varphi(x, v', v)u(x, v')\,\mathrm{d}v', \qquad \forall (x, v) \in \mathrm{D} \times \mathrm{V}, \\ u = 0, \qquad \text{on} \Gamma_-. \end{aligned} \tag{1.1}$$

This model is also called the radiative transfer equation when applied to uncharged particles other than neutrons (e.g. light and radiation). In the above formula and in the following, ∇ should be understood as the gradient with respect to the spatial coordinate x. Also, Γ_- and Γ_+ are the incoming/outgoing phase-space boundaries defined as

$$\begin{cases} \Gamma_- & := \{(x, v) \in \partial \mathrm{D} \times \mathrm{V} \mid v \cdot n(x) < 0\} \subset \partial \mathrm{D} \times \mathrm{V} \\ \Gamma_+ & := \{(x, v) \in \partial \mathrm{D} \times \mathrm{V} \mid v \cdot n(x) > 0\} \subset \partial \mathrm{D} \times \mathrm{V}, \end{cases} \tag{1.2}$$

where $n(x)$ denotes the unit outward normal at a point $x \in \partial \mathrm{D}$. The functions σ, κ and φ are nonnegative and they are the so-called optical parameters related to physical nuclear reactions. σ is the total cross-section, κ is the scattering cross-section and φ is associated to the fission cross-section.

It is well-known that under quite general assumptions (which we will recall, and discuss later on), it is possible to apply a Krein-Rutman theorem and show that the eigenvalue λ with smallest modulus is simple, real, and positive, and associated to a nonnegative and real eigenfunction u. In the following, we refer to this pair as the principal eigenpair, and we denote it by (u°, λ°). The value of λ° has a direct physical meaning since it describes the balance between the amount of fission effects on the one hand, and the amount of transport leakage and scattering effects on the other hand. The reactor is called supercritical if $\lambda^\circ < 1$ (the fission chain reaction escalates and needs to be slowed down), subcritical if $\lambda^\circ > 1$ (the chain reaction decreases and the energy production will eventually stop) and critical if $\lambda^\circ = 1$ (balanced chain reaction, with sustained energy production). Designing a reactor in which λ° is as close to 1 as possible is a key inverse problem in nuclear engineering. To reliably address it, it is necessary to be able to solve forward problems with rigorous, quantifiable accuracy. Developing principal strategies towards that end is the focus of this paper.

The proposed approach is iterative and involves, as a key constituent, the solution of intermediate source problems of the following form: for a given right-hand side function q, find u solution to

$$\begin{aligned} & v \cdot \nabla u(x, v) + \sigma(x, v) u(x, v) \\ & \quad - \int_{\mathrm{V}} \kappa(x, v', v) u(x, v') \, \mathrm{d}v' = q(x, v), \quad \forall (x, v) \in \mathrm{D} \times \mathrm{V}, \end{aligned} \tag{1.3}$$

endowed with some inflow boundary condition.

1.1.2 Central Goal

Of course, the prediction capability of numerical simulations based on such a model is challenged by a number of uncertainty sources, among them unavoidable model biases. However, once the model has been accepted as a basis for simulations, it is crucial to be able to guarantee the quality of the numerical outcome. The central objective of this paper is to contrive a conceptual pathway, along with the relevant analytical tools, towards rigorously quantifying the deviation of the numerical result from the exact, infinite-dimensional solution of the eigenvalue problem in a relevant norm. Note that this goes beyond asking for a priori error estimates, which have been the subject of extensive research in the past, and whose validity depends typically on extra regularity assumptions which are generally hard to check in practical applications. Instead, our approach hinges, in particular, on rigorous computable a posteriori quantities that hold without any excess regularity assumptions but just exploit suitable stable variational formulations of transport equations.

1.2 Prior Related Work and Challenges

The radiative transfer operator has been extensively studied from a diversity of point of views. From the perspective of nuclear reactor physics, we refer to [36, 43, 46] for a general overview. At the mathematical level, relevant results concerning the well-posedness and properties of the equation can be found in [1, 19, 38, 39]. There has also been an intense activity in numerical analysis and scientific computing in order to derive numerical solvers, and study their convergence properties. We refer to, e.g., [2, 29, 35, 44] for important results in this regard and to, e.g., [11, 12, 32, 41, 49] for further related relevant developments.

Until very recently, all these works follow the classical paradigm:

SA: *first discretize the continuous problem, then solve a fixed discrete problem.*

We refer to this as the "standard approach" (SA). While certainly this has its merits for many application scenarios, an accuracy controlled solution of eigenproblems for radiative transfer and Boltzmann type models poses particular challenges, revolving around the following issues: (a) the size of the discrete problem; (b) stability and solver accuracy; (c) discretization error uncertainty; (d) stability of the spectrum.

A few comments on these issues are in order. Regarding (a), the presence of two groups of variables (spatial and velocity variables) renders such problems high-dimensional. Therefore, when targeting relevant accuracy levels, related discretizations tend to become very large. This is aggravated by the fact that, lacking rigorous local error control, common approaches employ in essence uniform or quasi-uniform meshes that need to be fine enough to resolve relevant scales. In fact, such concerns are confirmed by the accuracy controlled source problem solver from [15] which is based on totally different concepts that will be taken up again later below. Matching the accuracy levels, achieved in [15] with the aid of adaptively refined grids, just using quasi-uniform meshes would have been infeasible. In addition, the appearance of global integral operators representing scattering and fission effects result in densely populated matrices. Hence the applicability of efficient sparse linear algebra techniques is limited.

As for (b), the lack of symmetry and of preconditioners of similar efficiency as for elliptic problems further impede on the efficient accurate solution of such very large scale problems unless strong model simplifications are accepted. Moreover, such discretizations are usually challenged by stability issues due to potentially strong transport effects and a lack of sufficient dissipation. Looking therefore for discretizations that have certain conservation properties still falls short of relating discrete solutions to the continuous one.

Concerning (c), when employing iterative solvers the question then is how accurately such large scale problems need to be solved and how to verify this accuracy because machine accuracy may be too demanding. It actually suffices if the accuracy of discrete solutions matches discretization error accuracy. Unfortunately, the actual discretization error remains largely uncertain. In fact, resorting to a priori error estimates as an orientation is not really a remedy, mainly for two reasons: first, they are typically derived under unrealistic regularity assumptions whose validity is hard to check in practice. Second, such error bounds involve norms of derivatives of the

unknown solution. In summary, an a priori quantification of the discretization error accuracy is in general not realistic.

This underlines the importance of deriving rigorous a posteriori error bounds and to understand how to use them to effectively reduce the error. In stark contrast to elliptic problems, sharp rigorous (lower and upper) a posteriori error bounds for kinetic models are far less developed. In particular, existing ones do not come with any comparable concise refinement strategy and corresponding convergence analysis. Early natural a posteriori bounds are based on stability in L_2, [37]. Underlying variational formulations for radiative transfer can be found in [3]. Mixed formulations based on parity principles are studied in [24]. An a posteriori error analysis for the graph norm based on duality principles is given in [30, 31] while in [40] a posteriori bounds are derived that are based on certain scaled graph norms. These latter works consider either constant kernels $\kappa(x, v, v')$ or kernels of the form $\kappa(x, v \cdot v')$. To the best of our knowledge these works do not provide a convergence analysis of concrete adaptive refinement methods based on such error bounds. While this has even been a long-standing open problem for elliptic problems, the non-locality of radiative transfer models seems to aggravate this issue. To the best of our knowledge, a first rigorous adaptive convergence theory for radiative transfer models has been developed in [15]. The critical new ingredient there is the combination of a posteriori error bounds, based on combining well-posedness in the sense of the Babuška–Nečas Theorem, with a quantifiable convergence of an idealized fixed point iteration in function space, see below for further comments in this regard.

Finally, as for (d), relying on rough estimates may be delicate because the continuity of spectra for such unsymmetric problems is less forgiving than for symmetric problems.

In summary, we claim that therefore a reliable accuracy quantification for problems of the present type is hard to realize on the basis of the standard approach (SA). In fact, to the best of our knowledge, [45] is the only previous deviation from (SA) for the criticality problem in the following sense. There, the authors study the convergence of an inexact inverse power iteration, viewed as a perturbation of such an iteration in function space. As explained in more detail in the subsequent section, this is close in spirit to the present work. However, the authors in [45] work under certain strong assumptions on the optical parameters: for the monoenergetic model[1] with uniformly constant scattering and fission functions κ and φ, the authors show that, on the continuous level, the general unsymmetric eigenproblem (1.1) is equivalent to a certain related eigenproblem for the scalar flux $\phi(x) = \int_V u(x, v)\mathrm{d}v$, involving a symmetric positive definite weakly singular integral operator (in space only). This correspondence to a symmetric problem (in a space of reduced dimension) allows them to give a convergence theory for an inexact inverse iteration. These techniques cannot be extended for nonconstant functions κ and φ. Moreover, concisely quantifying and certifying the necessary closeness of perturbed iterates to exact ones seems to be missing.

[1] In the monoenergetic model, the particles are assumed to have all the same kinetic energy so, for the analysis, one works with $|v| = 1$, thus $v \in \mathbb{S}^{d-1}$, the unit sphere of $\mathbb{R}^d$.

1.3 Conceptual Constituents and Novelty

In the light of the previous comments the approach proposed in the present paper deviates fundamentally from the standard approach (SA). A primary conceptual novelty for the present context lies in reversing the steps in (SA). In a nutshell, instead of discretizing right away the Boltzmann operator, we first contrive what we call an idealized or outer iteration in a suitable function space which needs to provably converge at a certain rate in the function space. The numerical method consists then of approximately carrying out each step in the idealized iteration within a step-dependent, dynamically updated accuracy tolerance. So, part of our task is to identify such tolerances that necessarily decrease in the course of the outer iteration (and which are expected to be rather coarse at initial stages). Hence, at no stage of the numerical solution process deals with a single (potentially large) discretization: each iteration step produces a discretization needed to stay sufficiently close to the idealized iterate.

The approach is reminiscent of the classical nested refinement idea, carried out, however, in an infinite-dimensional context. To obtain this way in the end results with certified accuracy, it is crucial that the target tolerance in each approximate outer iteration step is met. This underlines the role of a posteriori error bounds that need to be provided for this purpose. Such bounds should hold without any excess regularity assumptions and therefore need to rely on estimating infinite-dimensional residuals in the right norm. While such techniques have a long history in the context of elliptic problems this is not at all the case for transport dominated problems. Our approach to deriving such bounds is essentially based on stable ultraweak variational formulations for the involved transport equations. To our knowledge, this is a crucial distinction from all other computational approaches to the present eigenproblem. To aid further commenting on these ingredients we summarize this paradigm formally as follows:

(CC1) Contrive an idealized iteration in the relevant Hilbert space that can be shown to converge at a quantifiable rate.
(CC2) Determine perturbation tolerances within which the idealized iteration steps can be carried out so as to still warrant convergence to the exact continuous solution.
(CC3) Determine numerical tools that allow one to realize the tolerances in (CC2), certified by appropriate a posteriori error bounds.

Comments on (CC1)–(CC3) What is behind (CC1)–(CC3) and their relation to existing work? Regarding (CC1), we treat two such idealized iterations, namely Newton's method, and power iteration for different respective purposes. The obvious tempting aspect of Newton iterations is the potentially quadratic convergence, albeit generally under the condition that the initial guess is sufficiently close to the exact zero. Studying Newton's method in function space is certainly not a new idea, see e.g. [4]. Likewise, there is a wealth of literature addressing inexact Newton methods in Euclidean spaces, see e.g. [20, 48] as a few examples. Unfortunately, the findings we have been able to trace back do not apply to the issues arising in the specific context of (CC1)–(CC3). Here, it is essential to quantify suitable initial neighborhoods in conjunction with concrete bounds on resolvents when restricted to certain subspaces to end up with concrete bounds for the mapping properties of Jacobians as we get closer to the exact

principal eigenpair. In particular, we are not perturbing the operator but need to quantify the quality of approximate Newton updates depending on the state of the idealized iteration. Corresponding tools, presented in Sect. 3, are based on a refined analysis of the mapping properties of the resolvent. These preparations then allow us to address (CC2), namely to determine suitable perturbation bounds whose validity would still warrant convergence to the exact principal eigenpair at a quantifiable rate. Although this approach is closer in spirit to existing work on adaptive eigensolvers for elliptic problems (see e.g. [14, 16, 26]) the techniques in these works do not carry over to the present setting, among other things, due to the lack of symmetry in the present setting.

We further emphasize an essential distinction from working either in a fixed finite dimensional or in a fixed infinite-dimensional context which is typically the case in "non-adaptive" works. In our case the discretizations vary with evolving accuracy demands. While we show how to preserve even quadratic convergence for our perturbed idealized iteration, the numerical work will increase significantly because the tolerance will be seen to decrease also quadratically. Therefore, we show how properly chosen, less stringent tolerances lead to first order convergence where we can however choose the concrete error reduction rate. Which variant would be preferable in practice depends again on "critical problem parameters"—in the end on the spectral gap, namely the distance of the principal eigenvalue from the rest of the spectrum.

Finally, (CC3) brings in a central building block in our approach. An essential precondition is the fact that ideal iterations and their approximations are based on the same stable variational formulation. Specifically, we heavily exploit that, for the source problem (1.3), an accuracy controlled solver has been developed in [15]. Details about its implementation can be found in [15, 27]. It is based on the same paradigm of contriving first idealized iteration schemes—in this case a fixed point scheme—in function space followed by an approximate realization of each iteration step within dynamically updated accuracy tolerances, monitored by rigorous a posteriori quantities. This hinges crucially on uniformly stable variational formulations for the radiative transfer equation, that allows us to tightly estimate errors by residuals in the dual norm of the test space, despite the lack of coercivity. The evaluation of this dual norm, in turn, is facilitated by Discontinuous Petrov Galerkin (DPG) concepts. We briefly recapitulate in Sect. 2.2 the relevant facts needed in this paper. While rigorous a posteriori error control that does not require excess regularity is well-understood for elliptic problems and their close relatives, this is by far not the case for the Boltzmann operators arising in the present situation. Regarding the role of these latter findings for the present work, writing the source problem (1.3) abstractly as $\mathcal{B}u = q$, this allows us to build a routine

$$\mathcal{B},\ q,\ \eta \mapsto [\mathcal{B}^{-1}, q, \eta] \quad \text{such that} \quad \left\| \mathcal{B}^{-1} q - [\mathcal{B}^{-1}, q, \eta] \right\|_{L_2(\mathrm{D}\times\mathrm{V})} \leq \eta. \tag{1.4}$$

In this paper we invoke this routine as a central tool for an accuracy controlled solution of the criticality eigenvalue problem (1.1) which is the key objective of this paper. An important finding is that the eigensolver reduces in the end to a repeated application of the source problem solver subject to judiciously chosen tolerances.

Narrowing the information gap As the analysis of steps (CC1)–(CC2) shows, a final "perfectly certified" practical realization requires knowledge of certain "critical problem parameters" that enter for instance, the specification of perturbation tolerances. Our analysis shows that this can be reduced to knowing the spectral gap that quantifies the distance of the principal eigenvalue from the rest of the spectrum. Of course, this is in general not known, and an interesting question is whether an initial guess on the spectral gap can be upgraded computationally. The answer to this question is not clear. In fact, from an information theoretic point of view the present kind of spectral problem is known to be intrinsically hard, see e.g. [5]. So the present work may be viewed as addressing this fact from a "constructive angle". This is one of the issues addressed in Sect. 5.

But even if the spectral gap were known, a corresponding neighborhood of the principal eigenpair that guarantees convergence of the Newton scheme may be very small so that finding an admissible initial guess is far from trivial. In finite dimensions the power method often serves to complement a locally convergent higher order scheme because initial guesses are then far less constrained. According to (CC1)–(CC2), we discuss in Sect. 5 also the power method as an idealized iteration in function space. Since in this case a perturbed realization is simpler than for the Newton scheme, we only discuss convergence in function space, i.e., (CC1). Indeed, this is the critical issue in the present scenario of an unsymmetric compact operator. We have not been able to find any related results on quantifiable convergence in the literature. One obvious reason is that the typical proof strategies in the matrix case do not carry over. In particular, one cannot in general make use of an eigenbasis with quantifiable stability properties. Our approach is therefore quite different and hinges on functional calculus and Riesz projections. While one can establish convergence per se, a quantification of convergence properties seems to require at least some weak spectral decay properties. We show that if the operator belongs to a Schatten class the convergence of the power method can be quantified.

1.4 Outline

The paper is organized as follows. In Sect. 2 we present the weak formulation of source and eigenvalue problems (1.3) and (1.1). We detail our working assumptions on the optical parameters essentially in line with [1, 6, 15], and recall known results on the existence and uniqueness of the principal eigenpair (u°, λ°). Section 3 presents a Newton-based iterative scheme to approximate (u°, λ°). The scheme is formulated first in an idealized setting at an infinite-dimensional functional level for which we prove local quadratic convergence to (u°, λ°). Section 4 is then devoted to the approximate realization of this scheme along with the analysis of the perturbed scheme. In the spirit of [15], the strategy hinges on realizing the Newton updates within some judiciously chosen error tolerance, so as to arrive in the end at the desired target accuracy by controlling the deviation of the numerically perturbed iteration from the idealized one. In Sect. 4, we develop strategies that allow us to compute approximate Newton-updates within desired accuracy tolerances by just repeatedly applying the routine (1.4) in conjunction with accuracy controlled application schemes of global operators, also

discussed in [15]. Finally, in Sect. 5 we discuss the generation of a suitable initial guess that lies in the convergence neighborhood. We show that, under certain weak structural spectral assumptions, an inverse power iteration indeed converges. We conclude in Sect. 6 with some comments putting the findings in this paper into further perspective.

2 Weak Formulation of the Eigenvalue Problem

Our approach hinges on a stable ultra-weak formulation of the source problem (1.3). It requires certain (rather mild) assumptions on the optical parameters which we give in Sect. 2.1. We then define the formulation in Sect. 2.2 and discuss the properties of the eigenvalue with smallest modulus in Sect. 2.3.

2.1 Assumptions on the Domain and the Optical Parameters

We will work with the following assumptions which are in essence those adopted in related previous works like [1, 6, 15, 19, 28] as well:

(H1) The spatial domain D is bounded, convex and has piecewise $\mathcal{C}^1$ boundary. This guarantees that $\partial\,\mathrm{D}$ is smooth enough to have uniquely defined unit outward normals $n(x)$ at almost all points $x \in \partial\,\mathrm{D}$.

(H2) The velocity domain V is a compact subset of $\mathbb{R}^d$ which does not contain 0. Its d-dimensional Haar measure is assumed to have mass $|\,\mathrm{V}\,| = 1$. Moreover, V will always be assumed to contain a set which is homeomorphic to the sphere in $\mathbb{R}^d$. Any vector $v \in \mathbb{R}^d$ can be identified with the pair

$$(s, E) = (v/|v|, |v|^2/2) \in \mathbb{S}^{d-1} \times \mathbb{R}^+,$$

where s is the direction of propagation in the unit sphere $\mathbb{S}^{d-1}$ of $\mathbb{R}^d$, and E is the kinetic energy. Hence, we can adopt the identification $\mathrm{V} = \mathbb{S}^{d-1} \times \mathrm{E}$ with $\mathrm{E} = [E_{\min}, E_{\max}]$ and $0 < E_{\min} \leq E_{\max}$.

(H3) The nonnegative kernels $\kappa, \varphi : \mathrm{D} \times \mathrm{V} \times \mathrm{V} \to \mathbb{R}_+$ belong to $L_2(\mathrm{D} \times \mathrm{V} \times \mathrm{V})$ and satisfy

$$\begin{aligned} \max\left\{ \textstyle\int_{\mathrm{V}} \kappa(x, v', v)\mathrm{d}v,\ \int_{\mathrm{V}} \kappa(x, v, v')\mathrm{d}v \right\} &\leq M,\ (x, v') \in \mathrm{D} \times \mathrm{V} \\ \max\left\{ \textstyle\int_{\mathrm{V}} \varphi(x, v', v)\mathrm{d}v,\ \int_{\mathrm{V}} \varphi(x, v, v')\mathrm{d}v \right\} &\leq M,\ (x, v') \in \mathrm{D} \times \mathrm{V}, \end{aligned} \tag{2.1}$$

for some constant $M < \infty$.

(H4) The cross-section $\sigma : \mathrm{D} \times \mathrm{V} \to \mathbb{R}_+$ is bounded on $\mathrm{D} \times \mathrm{V}$ and there exists an $\alpha > 0$ such that

$$\begin{aligned} \min\Big\{ \sigma(x, v') - \int_{\mathrm{V}} \kappa(x, v', v)\mathrm{d}v,\ \sigma(x, v') \\ - \int_{\mathrm{V}} \kappa(x, v, v')\mathrm{d}v \Big\} \geq \alpha, \quad (x, v') \in \mathrm{D} \times \mathrm{V}\,. \end{aligned} \tag{2.2}$$

In the following, we call α the accretivity constant in view of property (2.22) below.

(H5) In addition we assume that the fission kernel is strictly positive, i.e., there exists a $c_f > 0$ such that

$$\varphi(x, v, v') \geq c_f, \quad (x, v, v') \in \mathrm{D} \times \mathrm{V} \times \mathrm{V}\,. \tag{2.3}$$

(H6) $\varphi \in C(\overline{\mathrm{D}}; L_2(\mathrm{D} \times \mathrm{V}))$.

Note that (H4) implies

$$\sigma(x, v) \geq \alpha > 0, \quad \forall\,(x, v) \in \mathrm{D} \times \mathrm{V}\,. \tag{2.4}$$

Moreover, (H4) and (H5) mean that the difference between absorption and scattering remains positive and that fission takes place everywhere in the phase space. It has been pointed out in [1] that these assumptions can be relaxed in favor of physically more realistic ones. In fact, at the expense of some additional technical effort, all conclusions remain valid when replacing (2.3) by the more physical assumption

$$\kappa(x, v, v') + \varphi(x, v, v') \geq c_f, \quad x \in \mathrm{D},\ v, v' \in \mathrm{V}, \tag{2.5}$$

where now the two kernels may vanish at mutually different places, see [6, Chapter II.2]. Under these premises the scattering and fission operators

$$\begin{aligned} u &\mapsto (\mathcal{K}u)(x, v) := \int_{\mathrm{V}} \kappa(x, v, v')u(x, v')\mathrm{d}v' \\ u &\mapsto (\mathcal{F}u)(x, v) := \int_{\mathrm{V}} \varphi(x, v, v')u(x, v')\mathrm{d}v' \end{aligned} \tag{2.6}$$

are bounded linear operators from $L_2(\mathrm{D} \times \mathrm{V})$ to $L_2(\mathrm{D} \times \mathrm{V})$, i.e.,

$$\mathcal{K},\ \mathcal{F} \in \mathcal{L}\big(L_2(\mathrm{D} \times \mathrm{V}), L_2(\mathrm{D} \times \mathrm{V})\big). \tag{2.7}$$

It will be convenient to introduce the bilinear forms

$$\begin{aligned} k(u, w) &:= \int_{\mathrm{D} \times \mathrm{V} \times \mathrm{V}} \kappa(x, v', v)u(x, v')w(x, v')\mathrm{d}x\mathrm{d}v\mathrm{d}v', \\ f(u, w) &:= \int_{\mathrm{D} \times \mathrm{V} \times \mathrm{V}} \varphi(x, v', v)u(x, v')w(x, v')\,\mathrm{d}x\mathrm{d}v\mathrm{d}v', \end{aligned} \tag{2.8}$$

corresponding to their variational definitions

$$(\mathcal{K}u)(w) = k(u, w), \quad (\mathcal{F}u)(w) = f(u, w), \quad u, w \in L_2(\mathrm{D} \times \mathrm{V}), \tag{2.9}$$

where $\mathcal{K}u$, $\mathcal{F}u$ are viewed as functionals acting on $L_2(\mathrm{D} \times \mathrm{V})$.

2.2 Variational Formulation of the Source Problem and Related Mapping Properties

In order to eventually arrive at a properly posed eigenproblem we will introduce first a stable weak formulation of the source problem (1.3) following in essence [15]. An important first step is to determine the mapping properties of the pure transport operator which, in strong form, is given as $\mathcal{T} u = v \cdot \nabla u + \sigma u$. Introducing the graph space

$$\begin{aligned} \mathbb{H}(\mathrm{D} \times \mathrm{V}) &:= \{w \in L_2(\mathrm{D} \times \mathrm{V}) : \ v \cdot \nabla w \in L_2(\mathrm{D} \times \mathrm{V})\} \\ &= \{u \in L_2(\mathrm{D} \times \mathrm{V}) : \mathcal{T} u \in L_2(\mathrm{D} \times \mathrm{V})\}, \end{aligned} \tag{2.10}$$

endowed with the norm

$$\|w\|^2_{\mathbb{H}(\mathrm{D} \times \mathrm{V})} := \|w\|^2_{L_2(\mathrm{D} \times \mathrm{V})} + \|v \cdot \nabla w\|^2_{L_2(\mathrm{D} \times \mathrm{V})}, \tag{2.11}$$

one readily sees that the bilinear form $t(u, w) := \int_{\mathrm{D} \times \mathrm{V}} w \mathcal{T} u \, dx dv$ is continuous over $\mathbb{H}(\mathrm{D} \times \mathrm{V}) \times L_2(\mathrm{D} \times \mathrm{V})$. Under the given assumptions on the optical parameters, one can show that $\mathcal{T}$ is lower bounded on $\mathbb{H}(\mathrm{D} \times \mathrm{V})$ (see e.g. [15]), i.e., there exists a positive constant $c > 0$ such that $\| \mathcal{T} u\|_{L_2(\mathrm{D} \times \mathrm{V})} \geq c \|u\|_{L_2(\mathrm{D} \times \mathrm{V})}$. Moreover, it is well-known that if a function $w \in \mathbb{H}(\mathrm{D} \times \mathrm{V})$ vanishes on either $\Gamma_\pm$ it possesses a trace on $\Gamma_\mp$ that belongs to the weighted L_2-space with norm $\|u\|^2_{L_{2,v}(\Gamma_\mp)} = \int_{\Gamma_\mp} n \cdot v u^2 \, d\gamma$ and

$$\mathbb{H}_{0,\mp}(\mathrm{D} \times \mathrm{V}) := \mathrm{clos}_{\|\cdot\|_{\mathbb{H}(\mathrm{D} \times \mathrm{V})}} \{u \in C^1(\mathrm{D} \times \mathrm{V}) : u|_{\Gamma_\mp} = 0\} \tag{2.12}$$

is a closed subspace of $\mathbb{H}(\mathrm{D} \times \mathrm{V})$. One then deduces that the operator $\mathcal{T}$, induced by the weak formulation

$$\langle \mathcal{T} u, w \rangle = \int_{\mathrm{D} \times \mathrm{V}} \big\{ v \cdot \nabla u + \sigma u \big\} w \, \mathrm{d}x \mathrm{d}v = \langle f, w \rangle, \quad w \in L_2(\mathrm{D} \times \mathrm{V}), \tag{2.13}$$

is an isomorphism from either space $\mathbb{H}_{0,\mp}(\mathrm{D} \times \mathrm{V})$ onto $L_2(\mathrm{D} \times \mathrm{V})$, and that the norm equivalences

$$\|u\|_{\mathbb{H}(\mathrm{D} \times \mathrm{V})} \eqsim \| \mathcal{T} u\|_{L_2(\mathrm{D} \times \mathrm{V})} \eqsim \|v \cdot \nabla u\|_{L_2(\mathrm{D} \times \mathrm{V})}, \quad u \in \mathbb{H}_{0,\mp}(\mathrm{D} \times \mathrm{V}), \tag{2.14}$$

hold, see e.g. [17]. Completely analogous relations are valid for the formal adjoint $\mathcal{T}^* u = -v \cdot \nabla u + \sigma u$ (to be distinguished from the dual operator $\mathcal{T}' : L_2(\mathrm{D} \times \mathrm{V}) \to \mathbb{H}_{0,\pm}(\mathrm{D} \times \mathrm{V})'$).

There is a second, referred to as ultraweak formulation. Applying integration by parts, yields

$$\langle \mathcal{T}\, u, w\rangle := \int_{\mathrm{D}\times\mathrm{V}} \big\{v\cdot\nabla u + \sigma u\big\} w \,\mathrm{d}x\mathrm{d}v = \int_{\mathrm{D}\times\mathrm{V}} u\big\{\sigma w - v\cdot\nabla w\big\}\mathrm{d}x\mathrm{d}v + \int_{\partial\,\mathrm{D}\times\mathrm{V}} (v\cdot n) u w d\gamma.$$

The integral over $\mathrm{D}\times\mathrm{V}$ is now well-defined for $u\in L_2(\mathrm{D}\times\mathrm{V})$ and $w\in\mathbb{H}(\mathrm{D}\times\mathrm{V})$. While $\int_{\partial\,\mathrm{D}\times\mathrm{V}}(v\cdot n)uwd\gamma$ is not defined for any $u\in L_2(\mathrm{D}\times\mathrm{V})$, substituting for u given boundary values $g\in L_{2,v}(\Gamma_-)$ and restricting the test functions w to $\mathbb{H}_{0,+}(\mathrm{D}\times\mathrm{V})$, the trace integral becomes $\int_{\Gamma_-}(v\cdot n)gwd\gamma$ and is thus well-defined. For homogeneous boundary conditions the boundary integral vanishes. The corresponding ultraweak formulation reads

$$\begin{aligned} t(u,w) &= \int_{\mathrm{D}\times\mathrm{V}} u\big\{\sigma w - v\cdot\nabla w\big\}\mathrm{d}x\mathrm{d}v = \langle u, \mathcal{T}^* w\rangle \\ &= \int_{\Gamma_-} |v\cdot n| g w \, d\gamma + \langle f, w\rangle, \quad w\in\mathbb{W} := \mathbb{H}_{0,+}(\mathrm{D}\times\mathrm{V}), \end{aligned} \tag{2.15}$$

induces the continuous extension of $\mathcal{T}:\mathbb{H}_{0,-}(\mathrm{D}\times\mathrm{V})\to L_2(\mathrm{D}\times\mathrm{V})$ to $\mathcal{T}: L_2(\mathrm{D}\times\mathrm{V})\to \big(H_{0,+}(\mathrm{D}\times\mathrm{V})\big)'$, where we use for convenience the same notation for the transport operator induced by the weak formulation $\langle \mathcal{T}\,u, w\rangle = \langle f, w\rangle$ or by (2.15). Note that in the latter case boundary conditions become part of the "right-hand side functional" and are thus natural boundary conditions. Since both formulations will be used later we summarize the relevant facts for later record as follows.

Theorem 2.1 *(see [15]) Let*

$$\mathbb{U} := L_2(\mathrm{D}\times\mathrm{V}), \quad \mathbb{W} := H_{0,+}(\mathrm{D}\times\mathrm{V}). \tag{2.16}$$

Under the assumptions (H1), (H2) and Eq. (2.4) of (H5), $\mathcal{T}$ defined by (2.13) is boundedly invertible as mapping from $H_{0,-}(\mathrm{D}\times\mathrm{V})$ onto $\mathbb{U}$, i.e.,

$$\mathcal{T}\in\mathcal{L}(H_{0,-}(\mathrm{D}\times\mathrm{V}),\mathbb{U}), \quad \mathcal{T}^{-1}\in\mathcal{L}(\mathbb{U}, H_{0,-}(\mathrm{D}\times\mathrm{V})). \tag{2.17}$$

Moreover, for $\mathcal{T}$, defined by (2.15), one has

$$\mathcal{T}\in\mathcal{L}(\mathbb{U},\mathbb{W}'), \quad \mathcal{T}^{-1}\in\mathcal{L}(\mathbb{W}',\mathbb{U}). \tag{2.18}$$

In very much the same way we can define the ultraweak formulation of the complete radiative transfer source problem as: find $u\in\mathbb{U}=L_2(\mathrm{D}\times\mathrm{V})$ such that for $q\in\mathbb{W}'\supset\mathbb{U}$

$$b(u,w) := \int_{\mathrm{D}\times\mathrm{V}} \Big\{ u\big(\sigma w - v\cdot\nabla w\big) - \int_{\mathrm{V}} \kappa\, u w \mathrm{d}v' \Big\} \mathrm{d}x\mathrm{d}v$$
$$= \int_{\Gamma_-} |v\cdot n| g w \, d\gamma + q(w), \quad w \in \mathbb{W}. \tag{2.19}$$

Note that, as natural boundary conditions, they need not be incorporated in any trial space. Of course, this part of the right-hand side disappears for homogeneous boundary conditions $g = 0$ considered in the eigenproblem.

In summary, for $q \in \mathbb{W}'$ finding $u \in \mathbb{U}$ such that

$$b(u,w) = q(w), \quad w \in \mathbb{W}, \tag{2.20}$$

is called ultra-weak formulation of the source problem (1.3) with homogeneous inflow boundary conditions. As usual, when $q \in L_2(\mathrm{D}\times\mathrm{V})$, we have $q(w) = \int_{\mathrm{D}\times\mathrm{V}} q w \mathrm{d}x\mathrm{d}v$. The accuracy controlled approximate solution of (2.20) will be a central constituent of subsequent eigensolver strategies.

Splitting

$$b(\cdot,\cdot) = t(\cdot,\cdot) - k(\cdot,\cdot),$$

where $t(\cdot,\cdot)$ is the bilinear form from (2.15) and $k(\cdot,\cdot)$ of the scattering part, it is straightforward to see that the bilinear forms $b(\cdot,\cdot), t(\cdot,\cdot), k(\cdot,\cdot), f(\cdot,\cdot)$ are continuous on $\mathbb{U}\times\mathbb{W}$. Defining in analogy to (2.9) the operators

$$(\mathcal{B}u)(w) = b(u,w), \quad (\mathcal{T}u)(w) = t(u,w), \quad \forall u \in \mathbb{U},\ w \in \mathbb{W},$$

we have, in line with the remarks at the end of the previous section, that $\mathcal{B}, \mathcal{T}, \mathcal{K}, \mathcal{F} \in \mathcal{L}(\mathbb{U}, \mathbb{W}')$. This allows us to interpret (2.20) as an operator equation

$$\mathcal{B}u = (\mathcal{T} - \mathcal{K})u = q. \tag{2.21}$$

Its unique solvability hinges on the mapping properties of the transport operator $\mathcal{T}$, given in Theorem 2.1. In particular, using the norm equivalences (2.14) in conjunction with the accretivity of the Boltzmann operator $\mathcal{B}$

$$(\mathcal{B}u, u) \geq \alpha \|u\|^2_{L_2(\mathrm{D}\times\mathrm{V})}, \quad \forall u \in \mathbb{H}_{0,-}(\mathrm{D}\times\mathrm{V}), \tag{2.22}$$

which readily follows from Eq. (2.2) of (H4), one derives the following facts with the aid of the Babuška–Nečas-Theory.

Theorem 2.2 *(see [15]) If hypothesis (H1)–(H4) hold, then $\mathcal{B}$ is a linear norm-isomorphism from $\mathbb{U}$ onto $\mathbb{W}'$, i.e., $\mathcal{B}^{-1} \in \mathcal{L}(\mathbb{W}', \mathbb{U})$ exists and the condition number $\kappa_{\mathbb{U},\mathbb{W}'}(\mathcal{B}) := \|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{W}')} \|\mathcal{B}^{-1}\|_{\mathcal{L}(\mathbb{W}',\mathbb{U})} < \infty$ is finite.*

Remark 2.3 The bound

$$\| \mathcal{B}^{-1} \|_{\mathcal{L}(\mathbb{U},\mathbb{U})} \leq \alpha^{-1}, \tag{2.23}$$

immediately follows from (2.22). Employing the norms $\| \mathcal{T}^* w\|_{\mathbb{U}}$ or $\| \mathcal{B}^* w\|_{\mathbb{U}}$ for $\mathbb{W}$ is expected to yield more favorable bounds for $\| \mathcal{B}^{-1} \|_{\mathcal{L}(\mathbb{W}',\mathbb{U})}$ when α is small in view of the Babuška–Nečas Theorem, see (3.18).

We proceed collecting a few more prerequisites for a proper formulation of the eigenproblem (1.1). A key observation concerns the weighted L_2-space

$$\mathbb{U}^{(\sigma)} := L^2_\sigma(\mathrm{D} \times \mathrm{V}), \quad \|u\|^2_{\mathbb{U}^{(\sigma)}} := \int_{\mathrm{D}\times\mathrm{V}} \sigma |u|^2 \mathrm{d}x\mathrm{d}v.$$

Under our assumptions on the total cross-section σ, one has the norm equivalence

$$\alpha \|u\|_{\mathbb{U}} \leq \|u\|_{\mathbb{U}^{(\sigma)}} \leq \|u\|_{\mathbb{U}} \max_{(x,v)\in \mathrm{D}\times\mathrm{V}} \sigma(x,v), \quad \forall u \in \mathbb{U}. \tag{2.24}$$

Therefore, the statements in Theorems 2.1, 2.2 remain valid (with constants depending now also on the lower and upper bound for σ). The following fact is shown in [6, Chapter II.2].

Lemma 2.4 *Adhering to the above notation, one has*

$$\| \mathcal{T}^{-1} \mathcal{K} \|_{\mathcal{L}(\mathbb{U}^{(\sigma)},\mathbb{U}^{(\sigma)})} \leq \rho < 1, \tag{2.25}$$

with ρ depending on α.

To ease accessibility of the underlying arguments we give a proof in Appendix A.

Theorem 2.5 *If (H1) to (H4) hold, then $\mathcal{K}$, $\mathcal{F}$, $\mathcal{T}^{-1}$ and $\mathcal{B}^{-1}$ are positive operators on*

$$\mathbb{U}^+ = L_2^+(\mathrm{D}\times\mathrm{V}) := \{u \in L_2(\mathrm{D}\times\mathrm{V}) : \ u(x,v) \geq 0, \ a.e. \ (x,v) \in \mathrm{D}\times\mathrm{V}\}.$$

Proof Since this is in essence a known fact (see e.g. [1, 6]) it suffices to sketch the arguments. First, $\mathcal{K}$ and $\mathcal{F}$ are positive by construction (see (2.6)), and positivity of $\mathcal{T}^{-1}$ follows from the method of characteristics. Now, to prove that $u = \mathcal{B}^{-1} q \in \mathbb{U}^+$ for every $q \in \mathbb{U}^+$, one readily checks that u is the unique fixed point

$$u = \mathcal{T}^{-1}(\mathcal{K} u + q).$$

Since, by (2.25), $\mathcal{T}^{-1}\mathcal{K}$ is a contraction on $\mathbb{U}^{(\sigma)}$, the fixed point iteration

$$u_{n+1} = \mathcal{T}^{-1}\mathcal{K} u_n + \mathcal{T}^{-1} q = \sum_{j=0}^{n} (\mathcal{T}^{-1}\mathcal{K})^j \mathcal{T}^{-1} q, \quad \forall n \geq 0, \tag{2.26}$$

converges to u in $\mathbb{U}^{(\sigma)}$ (hence in $\mathbb{U}$). Hence, we may write

$$u = \sum_{j=0}^{\infty} (\mathcal{T}^{-1}\mathcal{K})^j \mathcal{T}^{-1} q.$$

We conclude that $u \in \mathbb{U}^+$ from the fact that $\mathcal{K}$ and $\mathcal{T}^{-1}$ are positive operators. This concludes the proof that $\mathcal{B}^{-1}$ is a positive operator. □

The well-posedness of the operator Eq. (2.21), based on the above choices for $\mathbb{U}$ and $\mathbb{W}$, determines now the meaning of (1.1) as an operator eigenvalue problem:

$$\text{Find} \quad (u^\circ, \lambda^\circ) \in \tilde{\mathbb{U}} := \mathbb{U} \times \mathbb{R} \quad \text{such that} \quad \mathcal{B}u^\circ = \lambda^\circ \mathcal{F}u^\circ \tag{2.27}$$

through the weak formulation

$$b(u^\circ, w) = \lambda^\circ f(u^\circ, w), \quad \forall w \in \mathbb{W}. \tag{2.28}$$

The analysis of this eigenproblem is based on a reformulation involving the inverse of the Boltzmann operator $\mathcal{B}$ as explained in the next section. For a subsequent analysis, it will be useful to equip the space $\tilde{\mathbb{U}}$ with the norm

$$\|(u, \nu)\| := \left(\|u\|^2 + \nu^2\right)^{1/2}, \quad \forall (u, \nu) \in \tilde{\mathbb{U}}. \tag{2.29}$$

2.3 The Principal Eigenpair (u°, λ°)

The physical meaning of the principal eigenpair (u°, λ°) discussed in Sect. 1.1.1 suggests nonnegativity of u°, positivity and simplicity of λ°, and it is important that the neutronic model guarantees these properties. It is indeed folklore in the literature that, under certain assumptions on the optical parameters, the smallest eigenvalue in (2.27) is real, strictly positive, and simple. A detailed proof of these properties can be found in [6] (see also [1]). For the convenience of the reader, we sketch the underlying line of reasoning to an extent that is relevant for the subsequent developments in the present paper.

To that end, following in essence [6], a first step is to reformulate (2.27) such that Krein-Rutman theory is applicable. As shown in the previous section, $\mathcal{B}$ is an isomorphism from $\mathbb{U}^{(\sigma)}$ to $\mathbb{W}'$ so that

$$\mathcal{C} := \mathcal{B}^{-1}\mathcal{F} \in \mathcal{L}(\mathbb{U}^{(\sigma)}, \mathbb{U}^{(\sigma)}) = \mathcal{L}(\mathbb{U}, \mathbb{U}). \tag{2.30}$$

Thus, (2.27) is equivalent to finding $(u^\circ, \lambda^\circ) \in \tilde{\mathbb{U}}$ such that

$$\mathcal{C}u^\circ = (\lambda^\circ)^{-1} u^\circ. \tag{2.31}$$

Since we will now be concerned with operators in $\mathcal{L}(\mathbb{U},\mathbb{U}) = \mathcal{L}(\mathbb{U}^{(\sigma)},\mathbb{U}^{(\sigma)})$, we will simplify notation writing $\|\cdot\| = \|\cdot\|_{\mathbb{U}^{(\sigma)}} = \langle\cdot,\cdot\rangle^{1/2}$. For most statements, it does not matter whether which of the equivalent scalar products we are using except when the norm of $\mathcal{T}^{-1}\mathcal{K}$ matters. In this sense the above choice is in what follows just a convenient default. We sometimes refer to $\mathbb{U}^{(\sigma)}$ explicitly when the choice of the metric is important.

It will be convenient to denote for any $u \in \mathbb{U}$ by

$$\langle u\rangle := \mathrm{span}\{u\} \tag{2.32}$$

the linear span of u.

Theorem 2.6 *The operator $\mathcal{C} \in \mathcal{L}(\mathbb{U},\mathbb{U})$ is positive and compact. Moreover, there exists a unique simple largest positive eigenvalue μ° of*

$$\mathcal{C}\,u^\circ = \mu^\circ\,u^\circ. \tag{2.33}$$

The corresponding (up to normalization) unique eigenstate u° is non-negative. We henceforth refer to (μ°, u°) as principal eigenpair, and

$$\mathbb{U}_\circ := \langle u^\circ\rangle = \langle \mathcal{C}\,u^\circ\rangle \tag{2.34}$$

is the eigenspace associated to the principal eigenvalue λ° of our problem (2.27).

Proof We only sketch the main arguments of the proof following in essence [6]. For easier access to key arguments we present additional relevant details in Appendix B. Denoting $\mathcal{I}$ as the identity operator, we write

$$\mathcal{C} = (\mathcal{T}-\mathcal{K})^{-1}\,\mathcal{F} = (\mathcal{I}-\mathcal{T}^{-1}\mathcal{K})^{-1}\,\mathcal{T}^{-1}\,\mathcal{F}.$$

Next we use contractivity of $\mathcal{T}^{-1}\mathcal{K}$ on $\mathbb{U}^{(\sigma)}$ to write

$$\mathcal{C}\,u = \sum_{j=0}^{\infty}(\mathcal{T}^{-1}\mathcal{K})^j\big(\mathcal{T}^{-1}\mathcal{F}\big)u,$$

and use positivity of the operators $\mathcal{T}^{-1}, \mathcal{F}, \mathcal{K}$ to confirm positivity of $\mathcal{C}$.

Since $(\mathcal{I}-\mathcal{T}^{-1}\mathcal{K})^{-1} = \mathcal{B}^{-1} \in \mathcal{L}(\mathbb{W}',\mathbb{U}^{(\sigma)}) \supset \mathcal{L}(\mathbb{U}^{(\sigma)},\mathbb{U}^{(\sigma)})$ compactness of $\mathcal{C}$ follows as soon as we have established compactness of $\mathcal{T}^{-1}\mathcal{F}$. To that end, it suffices to confirm compactness of $\mathcal{F}^*\mathcal{T}^{-*} : \mathcal{L}(L_2(\mathrm{D}\times\mathrm{V}), L_2(\mathrm{D}\times\mathrm{V}))$. In view of (2.18), this in turn follows when $\mathcal{F}^* \in \mathcal{L}(H_{0,+}(\mathrm{D}\times\mathrm{V}), L_2(\mathrm{D}\times\mathrm{V}))$ is compact. Under the given assumptions on D, V, condition (H6) allows us to apply [19, Corollary 1, Appendix to §5, p. 415] which states exactly this fact. Since applying bounded operators to a compact one preserves compactness, the compactness of $\mathcal{C}$ follows from the above Neumann series. Krein-Rutman's Theorem asserts then that the spectral radius of $\mathcal{C}$ agrees with a positive eigenvalue μ° $(= \lambda^{\circ\,-1})$, associated with a non-negative eigenstate u°. The proof of strict positivity of u° (except on Γ_-) and of simplicity of

λ° can be found in the thesis [6, Chapter II.2]. We state these claims as a lemma whose proof is given for the interested reader in Appendix B.

□

Lemma 2.7 *The eigenvector λ° is simple. Moreover, the associated eigenstate u° is (up to normalization) unique and is strictly positive on* $(\mathrm{D} \times \mathrm{V}) \setminus \Gamma_-$.

This concludes the proof of Theorem 2.6.

In the original terms Theorem 2.6 can be restated as follows.

Corollary 2.8 *If (H1) to (H6) hold, then the eigenvalue problem* (2.27) *has a smallest simple real positive eigenvalue $\lambda^\circ = (\mu^\circ)^{-1}$ where μ° is the largest eigenvalue of $\mathcal{C}$ from Theorem 2.6. Its associated eigenvector u° is positive and equals the eigenstate of $\mathcal{C}$ associated to μ°.*

3 Strategy and Concepts for (CC1), (CC2)

This section is devoted to preparing the theoretical foundations for the paradigm outlined in Sect. 1.3. In view of the tempting prospect of a quadratic convergence we focus first on Netwon's method as idealized iteration in the infinite-dimensional space $\mathbb{U}$, see (CC1). We indeed show local quadratic convergence to the exact principal eigenpair (u°, λ°) of problem (1.1) provided that the initial guess belongs to a certain neighborhood that can be quantified in terms of the spectral gap (to be defined below) and mapping properties of resolvent. Based on these findings we analyze perturbed versions (CC2) that serve as the basis of numerical realizations. Such an envisaged numerical realization is to approximately realize the required operations within judiciously chosen, dynamically updated accuracy tolerances. Their validity needs to be confirmed in an a posteriori fashion to avoid requiring any excess regularity. A key role is played by establishing refined mapping properties of resolvents restricted to certain subspaces. Properties of resolvents for compact operators are most conveniently described for Hilbert spaces over $\mathbf{C}$. We therefore use in what follows for convenience the same notation $\mathbb{U}$ also for its complexification.

3.1 Newton's Method

As noted earlier, the problem of finding the principal eigenpair of the generalized operator eigenproblem (2.27) is equivalent to solving the standard eigenproblem (2.33) with $\mu^\circ = (\lambda^\circ)^{-1}$. We recall that we will use the notation $\|\cdot\| = \|\cdot\|_{\mathbb{U}^{(\sigma)}} =: \langle\cdot,\cdot\rangle^{1/2}$. It will be convenient to normalize eigenstates such that

$$\|\mathcal{C}\, u^\circ\| = 2 \quad \Leftrightarrow \quad \|u^\circ\| = 2\lambda^\circ = 2/\,\mu^\circ. \tag{3.1}$$

To formulate a Newton scheme, we define the residual function

$$\begin{aligned} &R : \tilde{\mathbb{U}} \to \tilde{\mathbb{U}} \\ &(u, \nu) \mapsto R(u, \nu) := \begin{bmatrix} R_1(u, \nu) \\ R_2(u, \nu) \end{bmatrix} = \begin{bmatrix} \mathcal{M}_\nu\, u \\ 1 - \|\mathcal{C}\, u\|^2/2. \end{bmatrix} \end{aligned}$$

where, for every $\nu \in \mathbf{C} \setminus \{0\}$,

$$\mathcal{M}_\nu := \mathcal{I} - \nu\, \mathcal{C}$$

is an operator from $\mathbb{U}$ into itself.

The (Fréchet) derivative $DR(\bar{u}, \bar{\nu}) : \tilde{\mathbb{U}} \to \tilde{\mathbb{U}}$ of R at a point $(\bar{u}, \bar{\nu})$, evaluated at a point (u, ν), is given by

$$DR(\bar{u}, \bar{\nu})(u, \nu) := \begin{bmatrix} \mathcal{M}_{\bar{\nu}} & -\mathcal{C}\, \bar{u} \\ -\langle \mathcal{C}\, \bar{u}, \cdot \rangle & 0 \end{bmatrix} \begin{bmatrix} u \\ \nu \end{bmatrix}. \tag{3.2}$$

Assuming for the moment invertibility of $DR(\bar{u}, \bar{\nu})(u, \nu)$, the Newton scheme consists in building a sequence

$$(u_{n+1}, \lambda_{n+1}) = (u_n + \delta_n^{(u)}, \lambda_n + \delta_n^{(\lambda)}), \quad \forall n \in \mathbb{N}, \tag{3.3}$$

starting from a given initial guess $(u_0, \lambda_0) \in \tilde{\mathbb{U}}$. The component $(\delta_n^{(u)}, \delta_n^{(\lambda)}) \in \tilde{\mathbb{U}}$ is an incremental update satisfying

$$DR(u_n, \lambda_n)(\delta_n^{(u)}, \delta_n^{(\lambda)}) = -R(u_n, \lambda_n). \tag{3.4}$$

The operator Eq. (3.4) has the block-structure of a saddle point problem which can be written as

$$\begin{bmatrix} \mathcal{M}_{\lambda_n} & -\mathcal{C}\, u_n \\ -\langle \mathcal{C}\, u_n, \cdot \rangle & 0 \end{bmatrix} \begin{bmatrix} \delta_n^{(u)} \\ \delta_n^{(\lambda)} \end{bmatrix} = -\begin{bmatrix} \mathcal{M}_{\lambda_n}\, u_n \\ 1 - \|\mathcal{C}\, u_n\|^2/2 \end{bmatrix}. \tag{3.5}$$

Assuming that $\lambda_n \neq \lambda^\circ$, and applying $\mathcal{M}_{\lambda_n}^{-1}$ to the first line yields

$$\delta_n^{(u)} = \delta_n^{(\lambda)}\, \mathcal{M}_{\lambda_n}^{-1}\, \mathcal{C}\, u_n - u_n. \tag{3.6}$$

Since, from the second line, $\left\langle \mathcal{C}\, u_n, \delta_n^{(u)} \right\rangle = 1 - \|\mathcal{C}\, u_n\|^2/2$, we deduce the value of $\delta_n^{(\lambda)}$ by taking the scalar product with $\mathcal{C}\, u_n$ in (3.6), which yields

$$\delta_n^{(\lambda)} = \frac{1 + \langle \mathcal{C}\, u_n, u_n \rangle - \|\mathcal{C}\, u_n\|^2/2}{\left\langle \mathcal{C}\, u_n, \mathcal{M}_{\lambda_n}^{-1}\, \mathcal{C}\, u_n \right\rangle}. \tag{3.7}$$

The updated iterates (3.3) can thus be written as

$$u_{n+1} = \delta_n^{(\lambda)}\, \mathcal{M}_{\lambda_n}^{-1}\, \mathcal{C}\, u_n = \frac{1 + \langle \mathcal{C}\, u_n, u_n\rangle - \|\mathcal{C}\, u_n\|^2/2}{\left\langle \mathcal{C}\, u_n, \mathcal{M}_{\lambda_n}^{-1}\, \mathcal{C}\, u_n \right\rangle}\, \mathcal{M}_{\lambda_n}^{-1}\, \mathcal{C}\, u_n \tag{3.8}$$

$$\lambda_{n+1} = \lambda_n + \frac{1 + \langle \mathcal{C}\, u_n, u_n\rangle - \|\mathcal{C}\, u_n\|^2/2}{\left\langle \mathcal{C}\, u_n, \mathcal{M}_{\lambda_n}^{-1}\, \mathcal{C}\, u_n \right\rangle}. \tag{3.9}$$

The whole scheme relies on evaluating $\mathcal{M}_{\lambda_n}^{-1}\, \mathcal{C}\, u_n$, i.e., on invertibility of $\mathcal{M}_{\lambda_n}$, which requires finding at each step the function $z_n \in \mathbb{U}$ such that

$$\mathcal{M}_{\lambda_n}\, z_n = \mathcal{C}\, u_n. \tag{3.10}$$

Solving (3.10) has to be handled with care since, on the one hand, the sequence $\big((u_n, \lambda_n)\big)_{n\geq 0}$ is to converge to the principal eigenpair (u°, λ°) but, on the other hand, as (u_n, λ_n) approaches (u°, λ°), the condition of the operator $\mathcal{M}_{\lambda_n} = \mathcal{I} - \lambda_n\, \mathcal{C}$ tends to infinity. In particular, we lose uniqueness since the operator $\mathcal{M}_{\lambda^\circ}$ has the nontrivial kernel $\mathbb{U}_\circ$ from (2.34). However, by uniqueness of the principal eigenpair (u°, λ°), the restriction $\mathcal{M}_{\lambda^\circ}|_{\mathbb{U}_\circ^\perp}$ of $\mathcal{M}_{\lambda^\circ}$ to $\mathbb{U}_\circ^\perp$ is injective. This motivates to study mapping properties at the principal eigenpair in order to rigorously prove a quantifiable convergence of the Newton scheme.

3.2 Some Prerequisites from the Spectral Theory of Compact Operators

To quantify the mapping properties of the Jacobian (3.2) as well as to prove later convergence of the power method as a means to generate suitable initial guesses (see Theorem 5.1 in Sect. 5) we draw on some classical facts from the spectral theory of compact operators on Hilbert spaces, see e.g. [25]. Let $\sigma(\mathcal{C})$ denote the spectrum of $\mathcal{C}$ (which contains $\{0\}$ for infinite-dimensional Hilbert spaces). Since there is hardly any confusion with indexing in the Newton iterates it will be convenient to enumerate the elements of the spectrum

$$\sigma(\mathcal{C}) = \{\mu_j : j = 1, \ldots, \infty\}$$

where $|\mu_j|$ decreases with increasing $j \in \mathbb{N}$. Thus, we replace the original notation (λ°, u°) for the principal eigenpair by (μ_1, u_1), keeping in mind that $\mu_1 = \mu^\circ = (\lambda^\circ)^{-1}$, $u^\circ = u_1$.

The open set $\rho(\mathcal{C}) := \mathbf{C} \setminus \sigma(\mathcal{C})$ is called resolvent set. The operator

$$\mathcal{R}_{\mathcal{C}}(\zeta) := (\zeta\, \mathcal{I} - \mathcal{C})^{-1}, \quad \zeta \in \rho(\mathcal{C}), \tag{3.11}$$

is called the resolvent operator which is obviously related to the operator $\mathcal{M}_\nu$ by

$$\mathcal{M}_\nu = \nu\, \mathcal{R}_{\mathcal{C}}(\nu^{-1})^{-1}. \tag{3.12}$$

$\mathcal{R}_{\mathcal{C}}(\zeta)$ is known to be an isomorphism on $\mathbb{U}$ for each $\zeta \in \rho(\mathcal{C})$.

Below as well as later in Sect. 5 the so called Riesz projections and related facts about functional calculus play a crucial role. For any subset $\omega \subset \sigma(\mathcal{C})$ they are defined as

$$\mathcal{E}_{\mathcal{C}}(\omega) = \frac{1}{2\pi i} \int_{\Gamma(\omega)} \mathcal{R}_{\mathcal{C}}(\zeta) d\zeta, \tag{3.13}$$

where we always assume a clockwise orientation of the closed rectifiable curve $\Gamma(\omega)$ forming the boundary of a domain $\Omega \subset \mathbf{C}$ that contains ω but does not intersect $\sigma(\mathcal{C}) \setminus \omega$, i.e.,

$$\Gamma(\omega) = \partial\Omega(\omega) \quad \text{and} \quad \omega \subset \Omega(\omega), \quad \Omega(\omega) \cap \sigma(\mathcal{C}) \setminus \omega = \emptyset.$$

$\mathcal{E}_{\mathcal{C}}(\omega)$ is independent of the specific contour Γ as long as Γ hast the above properties. Clearly, $\mathcal{C}$ commutes with $\mathcal{E}_{\mathcal{C}}(\omega)$. When $\omega = \{\mu\}$ contains only a single element $\mu \in \sigma(\mathcal{C}) \setminus \{0\}$ we write for convenience briefly $\mathcal{E}_{\mathcal{C}}(\mu) := \mathcal{E}_{\mathcal{C}}(\{\mu\})$.

Moreover, $\mathcal{E}_{\mathcal{C}}(\omega)$ is known to be a projection, i.e., $\mathcal{E}_{\mathcal{C}}(\Omega) = \mathcal{E}^2_{\mathcal{C}}(\Omega)$, and more generally

$$\mathcal{E}_{\mathcal{C}}(\mu)\mathcal{E}_{\mathcal{C}}(\mu') = \delta_{\mu,\mu'}\mathcal{E}_{\mathcal{C}}(\mu) \quad \forall\ \mu, \mu' \in \sigma(\mathcal{C}), \tag{3.14}$$

and for any $\omega \subset \sigma(\mathcal{C})$ one has a direct sum decomposition of $\mathbb{U}$

$$\mathbb{U} = \mathcal{E}_{\mathcal{C}}(\omega)\,\mathbb{U} \oplus \mathcal{E}_{\mathcal{C}}\big(\sigma(\mathcal{C}) \setminus \omega\big)\,\mathbb{U}, \tag{3.15}$$

into the invariant subspaces $\mathcal{E}_{\mathcal{C}}(\omega)\,\mathbb{U}$, $\mathcal{E}_{\mathcal{C}}\big(\sigma(\mathcal{C})\setminus\omega\big)\,\mathbb{U}$ of $\mathbb{U}$. Recall also, that for each $\mu \in \sigma(\mathcal{C})$ there exists a unique $n_\mu \in \mathbb{N}$ such that the spaces $V_\mu := \ker(\mu\mathcal{I} - \mathcal{C})^n$ agree for all $n \geq n_\mu$ and n_μ is the smallest number with this property. Moreover,

$$V_\mu := \mathcal{E}_{\mathcal{C}}(\mu)\,\mathbb{U}, \quad \mathcal{C}\,V_\mu \subseteq V_\mu. \tag{3.16}$$

In these terms the decomposition (3.15) can be refined as

$$\mathbb{U} = \bigoplus_{k\in\mathbb{N}} V_{\mu_k} = \bigoplus_{k\in\mathbb{N}} \mathcal{E}_{\mathcal{C}}(\mu_k)\,\mathbb{U}\,. \tag{3.17}$$

3.3 Mapping Properties at the Principal Eigenpair

In this section, we derive certain mapping properties of operators involving the principal eigenpair. These results are prerequisites for the convergence proof of the Newton scheme which is presented in the next section.

We begin with bounding $\mathcal{M}_\nu$. Recalling that $\mathcal{B}^{-1} \in \mathcal{L}(\mathbb{W}', \mathbb{U})$ and that $\mathbb{U}^{(\sigma)}$ is continuously embedded in $\mathbb{W}'$, it is useful to keep in mind that

$$\|\,\mathcal{C}\,\|_{\mathcal{L}(\mathbb{U}^{(\sigma)},\mathbb{U}^{(\sigma)})} \leq \|\,\mathcal{B}^{-1}\,\|_{\mathcal{L}(\mathbb{W}',\mathbb{U}^{(\sigma)})} \|\,\mathcal{F}\,\|_{\mathcal{L}(\mathbb{U}^{(\sigma)},\mathbb{W}')}. \tag{3.18}$$

A specification of this bound for $\|\mathcal{C}\|_{\mathcal{L}(\mathbb{U}^{(\sigma)},\mathbb{U}^{(\sigma)})}$ depends on the specific case at hand. We emphasize that such a bound may be much more favorable than the straightforward estimate $\|\mathcal{C}\|_{\mathcal{L}(\mathbb{U}^{(\sigma)},\mathbb{U}^{(\sigma)})} \le \|\mathcal{B}^{-1}\|_{\mathcal{L}(\mathbb{U}^{(\sigma)},\mathbb{U}^{(\sigma)})}\|\mathcal{F}\|_{\mathcal{L}(\mathbb{U}^{(\sigma)},\mathbb{U}^{(\sigma)})} \le \alpha^{-1}\|\mathcal{F}\|_{\mathcal{L}(\mathbb{U}^{(\sigma)},\mathbb{U}^{(\sigma)})}$ which may be too pessimistic for small α (see Remark 2.3). For convenience we abbreviate in what follows the norm of any operator $\mathcal{Z} \in \mathcal{L}(\mathbb{U}^{(\sigma)}, \mathbb{U}^{(\sigma)})$ as

$$\|\mathcal{Z}\| = \|\mathcal{Z}\|_{\mathcal{L}(\mathbb{U}^{(\sigma)},\mathbb{U}^{(\sigma)})} = \sup_{\|w\|=1} \|\mathcal{Z}w\|.$$

In subsequent discussions it will be convenient to use the shorthand notation

$$M_\nu := \|\mathcal{M}_\nu\| \le 1 + |\nu|\|\mathcal{C}\| \le 2\max\{1, |\nu|\|\mathcal{C}\|\}, \quad \forall \nu \in \mathbf{C}. \tag{3.19}$$

The following lemma serves as a major tool for analyzing the mapping properties of $\mathcal{M}_{\lambda^\circ}$.

Lemma 3.1 *Adhering to the above notation, one has*

$$\theta := \inf_{w\in\mathbb{U}_\circ^\perp} \sup_{v\in\mathbb{U}_\circ^\perp} \frac{\langle \mathcal{M}_{\lambda^\circ} w, v\rangle}{\|w\|\|v\|} > 0, \tag{3.20}$$

see (2.34) *for the definition of* $\mathbb{U}_\circ$.

Moreover, defining

$$\mathcal{M}^\circ_{\lambda^\circ} := P_{\mathbb{U}_\circ^\perp}\, \mathcal{M}_{\lambda^\circ}|_{\mathbb{U}_\circ^\perp} \in \mathcal{L}\big(\mathbb{U}_\circ^\perp, \mathbb{U}_\circ^\perp\big),$$

where $P_{\mathbb{U}_\circ^\perp}$ *denotes the* $\mathbb{U}$*-orthogonal projection of* $\mathbb{U}$ *to* $\mathbb{U}_\circ^\perp$*, we have*

$$\|(\mathcal{M}^\circ_{\lambda^\circ})^{-1}\|_{\mathcal{L}(\mathbb{U}_\circ^\perp,\mathbb{U}_\circ^\perp)} = \theta^{-1}. \tag{3.21}$$

Proof Since the range $\mathrm{ran}\,\big(\mathcal{M}_{\lambda^\circ}|_{\mathbb{U}_\circ^\perp}\big)$ is closed we need to show that $u^\circ \notin \mathrm{ran}\,\big(\mathcal{M}_{\lambda^\circ}|_{\mathbb{U}_\circ^\perp}\big)$. Suppose that there exists a $w_\circ \in \mathbb{U}_\circ^\perp$ such that $\mathcal{M}_{\lambda^\circ} w_\circ = u^\circ$. This means that

$$\langle u^\circ, v\rangle = \langle \mathcal{M}_{\lambda^\circ} w_\circ, v\rangle = \langle \mathcal{M}_{\lambda^\circ}(w_\circ + cu^\circ), v\rangle, \quad \forall\, c \in \mathbf{C},\ v \in \mathbb{U}\,.$$

This is the same as saying that there exists a $\hat{v} \in \mathbb{U}$ such that $\langle \mathcal{M}_{\lambda^\circ}\hat{v}, v\rangle = \langle u^\circ, v\rangle$ holds for all $v \in \mathbb{U}$. But, by the previous comments, $\hat{v}$ can also be written as $\hat{v} = \tilde{c}u^\circ + v_1$ for some $v_1 \in \mathbb{V} := \mathcal{E}_{\mathcal{C}}(\sigma(\mathcal{C})\backslash\{\mu^\circ\})\,\mathbb{U}$. Since $\mathcal{M}_{\lambda^\circ} u^\circ = 0$ this means that $\mathcal{M}_{\lambda^\circ} v_1 = u^\circ$ which is a contradiction since $\mathcal{M}_{\lambda^\circ} v_1 \in \mathbb{V}$.

To proceed, denote by $P_{\mathbb{U}_\circ^\perp}$ the $\mathbb{U}$-orthogonal projection of $\mathbb{U}$ to $\mathbb{U}_\circ^\perp$. We have shown above that

$$\ker\Big(P_{\mathbb{U}_\circ^\perp}(\mathcal{I} - \lambda^\circ\mathcal{C})|_{\mathbb{U}_\circ^\perp}\Big) = \{0\}. \tag{3.22}$$

Let

$$\mathcal{M}^\circ_{\lambda^\circ} := P_{\mathbb{U}^\perp_\circ}\, \mathcal{M}_{\lambda^\circ}|_{\mathbb{U}^\perp_\circ}$$

which, by definition, maps $\mathbb{U}^\perp_\circ$ into itself. As a closed subspace $\mathbb{U}^\perp_\circ$ is also a Hilbert space endowed with the norm $\|\cdot\|$. Moreover, $\mathcal{M}^\circ_{\lambda^\circ} = \mathcal{I}|_{\mathbb{U}^\perp_\circ} - \lambda^\circ P_{\mathbb{U}^\perp_\circ}\mathcal{C}|_{\mathbb{U}^\perp_\circ}$ and $P_{\mathbb{U}^\perp_\circ}\mathcal{C}$ is a compact operator taking $\mathbb{U}^\perp_\circ$ into itself. Hence Fredholm's alternative is valid and says, on account of (3.22), that $\mathcal{M}^\circ_{\lambda^\circ}$ maps $\mathbb{U}^\perp_\circ$ onto itself. By the Open Mapping Theorem, $\mathcal{M}^\circ_{\lambda^\circ} : \mathbb{U}^\perp_\circ \to \mathbb{U}^\perp_\circ$ is boundedly invertible. The Babuška–Nečas Theorem then says that

$$\theta = \inf_{w\in\mathbb{U}^\perp_\circ} \sup_{v\in\mathbb{U}^\perp_\circ} \frac{\langle \mathcal{M}^\circ_{\lambda^\circ} w, v\rangle}{\|w\|\|v\|} > 0.$$

Since $\langle \mathcal{M}^\circ_{\lambda^\circ} w, v\rangle = \langle \mathcal{M}_{\lambda^\circ} w, v\rangle$ for all $v \in \mathbb{U}^\perp_\circ$ the assertion follows. □

The quantitative mapping properties of the Jacobian (3.2), and hence the convergence properties of the Newton scheme, rely in essence on the size of θ as shown later below. The convergence of the power iteration, discussed in Sect. 5, instead depends on another spectral property of $\mathcal{C}$ that we introduce next. Let

$$\mathfrak{q} = \mathfrak{q}_\mathcal{C} := \frac{\lambda^\circ}{\Lambda} < 1, \quad \text{where} \quad \Lambda := \max\big\{|\lambda| : \lambda \neq \lambda^\circ,\ \ker \mathcal{M}_\lambda \neq \{0\}\big\}, \tag{3.23}$$

which we use to encode the (relative) spectral gap

$$\Delta := 1 - \mathfrak{q}. \tag{3.24}$$

Although we will not make any direct use of the following remarks we pause to briefly comment on the relation between θ and Δ which is in general a strict lower bound for

$$\overline{\Delta} := \Big|1 - \frac{\lambda^\circ}{\lambda_\Lambda}\Big| = \Big|1 - \frac{\lambda^\circ}{\bar\lambda_\Lambda}\Big|, \tag{3.25}$$

where λ_Λ is the eigenvalue with $|\lambda_\Lambda| = \Lambda$. More precisely, $\Delta = \overline{\Delta}$ if and only if $\lambda_\Lambda \in \mathbb{R}$ which is, for instance, the case when $\mathcal{C}$ is a normal operator.

In fact, it is $\overline{\Delta}$ that relates more closely to θ as explained next. By (3.21), θ is the smallest singular value of the mapping $\mathcal{M}^\circ_{\lambda^\circ}$ and hence of its adjoint $(\mathcal{M}^\circ_{\lambda^\circ})^* \in \mathcal{L}(\mathbb{U}^\perp_\circ, \mathbb{U}^\perp_\circ)$. Let us denote by $r_\circ \in \mathbb{U}^\perp_\circ$ the associated left singular vector of $\mathcal{M}^\circ_{\lambda^\circ}$ with the standard normalization $\|r_\circ\| = 1$. Let λ_Λ be a generalized eigenvalue of $\lambda\,\mathcal{C}\,u = u$ with second-smallest modulus, i.e., $|\lambda_\Lambda| = \Lambda$. Then $\bar\lambda_\Lambda$ is a generalized eigenvalue of the adjoint problem $\nu\,\mathcal{C}^*\,u = u$ (with the same modulus) and let $u^*_\Lambda \in \mathbb{U}^\perp_\circ$ denote the associated eigenstate.

Remark 3.2 Adhering to the above notation and denoting by $\mathbb{U}_\circ^{*\perp} := \langle u_\circ^* \rangle^\perp$, where $u_\circ^*$ is the principal eigenstate of $\mathcal{C}^*$, we have

$$\mathbb{V}^* := \mathcal{E}_{\mathcal{C}^*}(\sigma(\mathcal{C}^*) \setminus \{\mu^\circ\}) = \mathbb{U}_\circ^\perp, \quad \mathbb{V} = \mathcal{E}_{\mathcal{C}}(\sigma(\mathcal{C}) \setminus \{\mu^\circ\}) = \mathbb{U}_\circ^{*\perp}.$$

(see (C2) in Appendix C), and

$$\Delta |\langle r_\circ, u_\Lambda^* \rangle| \le \overline{\Delta} |\langle r_\circ, u_\Lambda^* \rangle| \le \theta \le \Big(1 - \operatorname{dist}\big(\mathbb{U}_\circ^\perp, \mathbb{U}_\circ^{*\perp}\big)\Big)\, \overline{\Delta} \tag{3.26}$$

where

$$\operatorname{dist}\big(\mathbb{U}_\circ^\perp, \mathbb{U}_\circ^{*\perp}\big) := \max\Big\{ \max_{\substack{w \in \mathbb{U}_\circ^\perp \\ \|w\|=1}} \|w - P_{\mathbb{U}_\circ^{*\perp}} w\|,\ \max_{\substack{w \in \mathbb{U}_\circ^{*\perp} \\ \|w\|=1}} \|w - P_{\mathbb{U}_\circ^{\perp}} w\| \Big\}.$$

Moreover, when $\mathcal{C}$ is a normal operator we have

$$\theta = \overline{\Delta} = \Delta. \tag{3.27}$$

The discrepancies in (3.26) can be viewed as quantifying the deviation of $\mathcal{C}$ from normality.

We provide details behind the above claims in Appendix C.

The next theorem shows that the mapping $DR(u^\circ, \lambda^\circ) \in \mathcal{L}(\tilde{\mathbb{U}}, \tilde{\mathbb{U}})$ is boundedly invertible.

Theorem 3.3 *At the principal eigenpair (u°, λ°), the mapping $DR(u^\circ, \lambda^\circ) \in \mathcal{L}(\tilde{\mathbb{U}}, \tilde{\mathbb{U}})$ is boundedly invertible so that, in particular, there exists a positive constant β such that*

$$\|DR(u^\circ, \lambda^\circ)^{-1}\| := \|DR(u^\circ, \lambda^\circ)^{-1}\|_{\mathcal{L}(\tilde{\mathbb{U}}, \tilde{\mathbb{U}})} \le \beta, \tag{3.28}$$

where

$$\beta \le \left(\left(\frac{1}{\theta} + \frac{1}{2}\Big(1 + \frac{M_{\lambda^\circ}}{\theta}\Big) \right)^2 + \frac{1}{4}\Big(1 + \frac{M_{\lambda^\circ}}{\theta}\Big)^2 \Big(1 + \frac{M_{\lambda^\circ}}{2}\Big)^2 \right)^{1/2}. \tag{3.29}$$

Proof Recall that

$$DR(u^\circ, \lambda^\circ) = \begin{bmatrix} \mathcal{M}_{\lambda^\circ} & -\mathcal{C}\, u^\circ \\ -\langle \mathcal{C}\, u^\circ, \cdot \rangle & 0 \end{bmatrix}$$

and define

$$\begin{cases} a(u, z) & := \langle \mathcal{M}_{\lambda^\circ} u, z \rangle, \quad \forall (u, z) \in \mathbb{U} \times \mathbb{U}, \\ b(u, \nu) & := -\nu \langle \mathcal{C}\, u^\circ, u \rangle, \quad \forall (u, \nu) \in \mathbb{U} \times \mathbb{R}. \end{cases} \tag{3.30}$$

Then, for any given $(g, \zeta) \in \tilde{\mathbb{U}}$, the weak formulation of finding $(u, \nu) \in \tilde{\mathbb{U}}$ such that $DR(u^\circ, \lambda^\circ)(u, \nu) = (g, \zeta)$, is given by

$$\begin{aligned} a(u, z) + b(z, \nu) &= \langle g, z \rangle, \quad \forall z \in \mathbb{U} \\ b(u, \alpha) \qquad\quad &= \alpha\zeta, \qquad \forall \alpha \in \mathbb{R}. \end{aligned} \tag{3.31}$$

We show next that (3.31) satifies the Ladyzhenskaya–Babu šska–Brezzi (LBB)-conditions (see Theorem 2.34 of [23]). We have already seen that both bilinear forms are continuous on $\mathbb{U} \times \mathbb{U}$ and $\mathbb{U} \times \mathbb{R}$, respectively. Specifically, from (3.19) and (3.1), it holds that

$$\|a\| := \sup_{u\in\mathbb{U}} \sup_{z\in\mathbb{U}} \frac{a(u, z)}{\|u\|\|z\|} \le M_{\lambda^\circ}, \;\; \|b\| := \sup_{\nu\in\mathbb{R}} \sup_{z\in\mathbb{U}_\circ} \frac{|b(z, \nu)|}{\|z\||\nu|} \le \|\mathcal{C}\, u^\circ\| = 2. \tag{3.32}$$

Next, we observe that we have

$$\mathbb{U}_\circ^\perp = \{u \in \mathbb{U} \;:\; b(u, \nu) = 0, \; \forall \nu \in \mathbb{R}\},$$

so that, by (3.20) in Lemma 3.1,

$$\inf_{u\in\mathbb{U}_\circ^\perp} \sup_{z\in\mathbb{U}_\circ^\perp} \frac{a(u, z)}{\|u\|\|z\|} \ge \theta.$$

Finally,

$$\inf_{\nu\in\mathbb{R}} \sup_{z\in\mathbb{U}} \frac{|b(z, \nu)|}{\|z\||\nu|} = \inf_{\nu\in\mathbb{R}} \sup_{z\in\mathbb{U}} \frac{|\nu|\,\langle \mathcal{C}\, u^\circ, z\rangle}{\|z\||\nu|} = \|\mathcal{C}\, u^\circ\| = 2,$$

where we have again used the normalization $\|\mathcal{C}\, u^\circ\| = 2$ from (3.1). The saddle point problem (3.31) is thus well-posed and we have the a priori estimates

$$\|u\| \le \frac{1}{\theta}\|g\| + \frac{1}{2}\Big(1 + \frac{M_{\lambda^\circ}}{\theta}\Big)|\zeta|, \quad |\nu| \le \frac{1}{2}\Big(1 + \frac{M_{\lambda^\circ}}{\theta}\Big)\|g\| + \frac{M_{\lambda^\circ}}{4}\Big(1 + \frac{M_{\lambda^\circ}}{\theta}\Big)|\zeta|. \tag{3.33}$$

Using the norm (2.29) for the space $\tilde{\mathbb{U}}$, the claims (3.28), (3.29) follow by straightforward calculations.

□

3.4 Perturbation Results

Recall from (3.3) and (3.4), that each Newton iteration with current approximation (u_n, λ_n) to the principal eigenpair (u°, λ°) requires inverting the operator $DR(u_n, \lambda_n)$. The main result in this section states that $DR(u, \nu)$ has indeed a uniformly bounded

inverse for all (u, v) in a full neighborhood of (u°, λ°). Thus inverting $DR(u_n, \lambda_n)$ as required in (3.4) is allowed as long as the (u_n, λ_n) remain in that neighborhood. Proving this statement requires extending Theorem 3.3 to a neighborhood of (u°, λ°). We present this result in our next theorem. To introduce it, for a given metric space X, it will be convenient to denote by $B(u, \tau)$ the closed ball with center $u \in X$ and radius $\tau \geq 0$.

Theorem 3.4 *There exists a radius $\tau > 0$ that depends on $\theta, \lambda^\circ, \|\mathcal{C}\|$ such that*

$$\|DR(u, v)^{-1}\| \leq \bar{\beta}, \quad \forall\, (u, v) \in \mathcal{N}, \tag{3.34}$$

where

$$\mathcal{N} := B(u^\circ, \tau) \times B\Big(\lambda^\circ, \frac{\theta}{4\|\mathcal{C}\|}\Big). \tag{3.35}$$

The constant $\bar{\beta}$ depends only on $\lambda^\circ, \theta, \|\mathcal{C}\|$, and it holds that

$$\begin{aligned} 0 < \bar{\beta} \leq & \left(\frac{16}{\theta^2} + \Big(\frac{2\lambda^\circ}{2\lambda^\circ - \tau}\Big)^2 \left(1 + 4\frac{\bar{M}}{\theta}\right)^2\right)^{1/2} \\ & + \frac{2\lambda^\circ}{2\lambda^\circ - \tau}\left(1 + 4\frac{\bar{M}}{\theta}\right)\left(1 + \frac{\bar{M} 4\lambda^{\circ 2}}{2\lambda^\circ - \tau)^2}\right)^{1/2}, \end{aligned} \tag{3.36}$$

where

$$\tau = \min\left\{\frac{1}{8\|\mathcal{C}\|}, \frac{2\theta}{25(1 + \lambda^\circ\|\mathcal{C}\| + \theta/4)\|\mathcal{C}\|}, \frac{\theta}{25\bar{M}}\right\}, \qquad \bar{M} := 1 + \lambda^\circ\|\mathcal{C}\| + \frac{\theta}{4}. \tag{3.37}$$

Remark 3.5 Notice that the admissible radius τ depends on θ (even linearly so, when replacing the middle expression in the definition of τ by the bound $\frac{2\theta}{25(1+\lambda^\circ\|\mathcal{C}\|)\|\mathcal{C}\|}$) which, in view of (3.26), reflects an increasing difficulty when the spectral gap Δ is small. Specifically, suppose that $\|\mathcal{C}\|$ has moderate size of order one, then $\bar{\beta} \lesssim \theta^{-1} \max\{1, \lambda^\circ\}$.

The remainder of this section is devoted to the proof of Theorem 3.4, based on a number of perturbation results.

Lemma 3.6 *For all $v \in B(\lambda^\circ, \frac{\theta}{2\|\mathcal{C}\|})$, $\mathcal{M}_v$ is boundedly invertible on $\mathbb{U}_\circ^\perp$, and*

$$\inf_{w \in \mathbb{U}_\circ^\perp} \sup_{v \in \mathbb{U}_\circ^\perp} \frac{\langle \mathcal{M}_v\, w, v\rangle}{\|w\|\|v\|} \geq \frac{\theta}{2}. \tag{3.38}$$

Proof Boundedness of $\mathcal{M}_\nu$ follows from (3.19). Regarding the inf-sup condition, we have for every $w, v \in \mathbb{U}_\circ^\perp$,

$$\begin{aligned}\langle \mathcal{M}_\nu\, w, v\rangle &= \langle \mathcal{M}_{\lambda^\circ}\, w, v\rangle - \big\langle (\nu-\lambda^\circ)\,\mathcal{C})w, v\big\rangle \geq \theta\|w\|\|v\| - (|\nu|-\lambda^\circ)|\,\langle \mathcal{C}\, w, v\rangle \\ &\geq (\theta - |\nu-\lambda^\circ|\,\|\,\mathcal{C}\,\|)\|u\|\|v\|\end{aligned}$$

which confirms the claim. The same argument shows validity of the swapped inf-sup condition

$$\inf_{v\in\mathbb{U}_\circ^\perp}\sup_{w\in\mathbb{U}_\circ^\perp}\frac{\langle \mathcal{M}_\nu\, w, v\rangle}{\|w\|\|v\|} \geq \frac{\theta}{2}.$$

The claim follows now from the Babuška–Nečas Theorem [23]. □

Lemma 3.7 *For every $\varepsilon > 0$,*

$$\langle \mathcal{C}\, u, u\rangle \geq a_\varepsilon \|u\|^2, \quad \forall u \in B(u^\circ, \varepsilon\|u^\circ\|),$$

where

$$a_\varepsilon := (1-\varepsilon)^2\big((\lambda^\circ)^{-1} - \varepsilon\|\,\mathcal{C}\,\|(2+\varepsilon)\big). \tag{3.39}$$

Moreover, there exists $\varepsilon_0 = \varepsilon_0(\|\,\mathcal{C}\,\|, \lambda^\circ)$ such that

$$a_\varepsilon \geq \frac{1}{2\lambda^\circ}, \quad \forall \varepsilon \leq \varepsilon_0, \tag{3.40}$$

where under the assumption $\varepsilon \leq 1$

$$\varepsilon_0 = \frac{1}{16\|\,\mathcal{C}\,\|\lambda^\circ} \tag{3.41}$$

suffices to ensure (3.40).

Proof Using that $\langle \mathcal{C}\, u^\circ, u^\circ\rangle = (\lambda^\circ)^{-1}\|u^\circ\|^2$, we have for any u such that $\|u-u^\circ\| \leq \varepsilon\|u^\circ\|$,

$$\begin{aligned}\langle \mathcal{C}\, u, u\rangle &= \big\langle \mathcal{C}(u-u^\circ), u-u^\circ\big\rangle + \big\langle \mathcal{C}\, u^\circ, u^\circ\big\rangle + \big\langle \mathcal{C}(u-u^\circ), u^\circ\big\rangle + \big\langle u-u^\circ, \mathcal{C}\, u^\circ\big\rangle \\ &\geq (\lambda^\circ)^{-1}\|u^\circ\|^2 - \|\,\mathcal{C}\,\|\|u-u^\circ\|^2 - 2\|\,\mathcal{C}\,\|\|u^\circ\|\,\|u-u^\circ\| \\ &\geq (\lambda^\circ)^{-1}\|u^\circ\|^2 - \|\,\mathcal{C}\,\|\varepsilon^2\|u^\circ\|^2 - 2\varepsilon\|\,\mathcal{C}\,\|\|u^\circ\|^2 = \big((\lambda^\circ)^{-1} - \varepsilon\|\,\mathcal{C}\,\|(2+\varepsilon)\big)\|u\|^2 \\ &\geq (1-\varepsilon)^2\big((\lambda^\circ)^{-1} - \varepsilon\|\,\mathcal{C}\,\|(2+\varepsilon)\big)\|u\|^2.\end{aligned}$$

□

The following observation extends Lemma 3.1 to pairs (u, ν) near (u°, λ°).

Lemma 3.8 *Let* $(u, v) \in B(u^\circ, \varepsilon\|u^\circ\|) \times B\left(\lambda^\circ, \frac{\theta}{4\|\mathcal{C}\|}\right)$. *Then*

$$\left\langle \mathcal{M}_v\, z, z' \right\rangle \geq \frac{\theta}{4}\|z\|\|z'\|, \quad \forall\, z, z' \in \operatorname{span}\{\mathcal{C}\, u\}^\perp \tag{3.42}$$

provided that $\varepsilon \leq \varepsilon_1$, *where* $\varepsilon_1 = \varepsilon_1(\|\mathcal{C}\|, \lambda^\circ, \theta)$ *is defined in* (3.50) *below.*

Proof Note first that the unperturbed result $u = u^\circ$ and $v = \lambda^\circ$ was proven in Lemma 3.1. To prove the general case, we fix $\varepsilon > 0$ and let $(u, v) \in B(u^\circ, \varepsilon\|u^\circ\|) \times B(\lambda^\circ, \frac{\theta}{4\|\mathcal{C}\|})$. That is,

$$\|u - u^\circ\| \leq \varepsilon\|u^\circ\| \quad \text{and} \quad |v - \lambda^\circ| \leq \frac{\theta}{4\|\mathcal{C}\|},$$

i.e., u and u° differ by at most ε in relative accuracy (and similarly for v and λ°).

For the fixed u, let now $z \in \operatorname{span}\{\mathcal{C}\, u\}^\perp$, and let $z_\circ^\perp := \mathcal{P}_{\mathbb{U}_\circ^\perp}(z)$ be its projection to $\mathbb{U}_\circ^\perp$. The first step of the proof consists in showing that z and $z_\circ^\perp$ also differ by the order of ε in relative accuracy. Denoting by $\mathcal{P}_v\, w = \langle w, v\rangle v/\|v\|^2$ the orthogonal projection of an element $w \in \mathbb{U}$ to the subspace spanned by a function $v \in \mathbb{U}$, we have

$$z - z_\circ^\perp = \mathcal{P}_{u^\circ}\, z = \mathcal{P}_{\mathcal{C}\, u^\circ}\, z = \frac{\langle z, \mathcal{C}\, u^\circ\rangle\, \mathcal{C}\, u^\circ}{\|\mathcal{C}\, u^\circ\|^2} = \frac{\langle z, \mathcal{C}(u^\circ - u)\rangle\, \mathcal{C}\, u^\circ}{\|\mathcal{C}\, u^\circ\|^2}, \tag{3.43}$$

where we have used that $\langle z, \mathcal{C}\, u\rangle = 0$. Taking norms, yields

$$\|z - z_\circ^\perp\| \leq \|z\|\|\mathcal{C}\|\frac{\|u - u^\circ\|}{\|\mathcal{C}\, u^\circ\|} \leq \bar{C}\varepsilon\|z\|, \tag{3.44}$$

with

$$\bar{C} = \lambda^\circ\|\mathcal{C}\|. \tag{3.45}$$

Note next that, using (3.19), the one-parameter family of linear operators $\mathcal{M}_v$ can be bounded uniformly for $v \in B\left(\lambda^\circ, \frac{\theta}{4\|\mathcal{C}\|}\right)$ as

$$M_v = \|\mathcal{M}_v\| \leq \max_{|v-\lambda^\circ| \leq \frac{\theta}{4\|\mathcal{C}\|}} \left\{1 + |v|\|\mathcal{C}\|\right\} \leq 1 + \lambda^\circ\|\mathcal{C}\| + \frac{\theta}{4} =: \bar{M}. \tag{3.46}$$

We abbreviate $\tilde{z}_\circ^\perp := \mathcal{P}_{u^\circ}\, \tilde{z}$ Under the given assumption on v we can invoke now Lemma 3.6 and (3.44) to conclude that for any $z, \tilde{z} \in \operatorname{span}\{\mathcal{C}\, u\}^\perp$,

$$\begin{aligned}\langle \mathcal{M}_\nu\, z, \tilde{z}\rangle &= \left\langle \mathcal{M}_\nu(z - z_\circ^\perp), (\tilde{z} - \tilde{z}_\circ^\perp)\right\rangle + \left\langle \mathcal{M}_\nu\, z_\circ^\perp, \tilde{z}_\circ^\perp\right\rangle + \left\langle \mathcal{M}_\nu\, z_\circ^\perp, (\tilde{z} - \tilde{z}_\circ^\perp)\right\rangle \\ &\quad + \left\langle \mathcal{M}_\nu(z - z_\circ^\perp), \tilde{z}_\circ^\perp\right\rangle \\ &\geq \frac{\theta}{2}\|z_\circ^\perp\|\|\tilde{z}_\circ^\perp\| - M_\nu\|z - z_\circ^\perp\|\|\tilde{z} - \tilde{z}_\circ^\perp\| \\ &\quad - M_\nu(\|\tilde{z}_\circ^\perp\|\|z - z_\circ^\perp\| + \|z_\circ^\perp\|\|\tilde{z} - \tilde{z}_\circ^\perp\|) \\ &\geq \frac{\theta}{2}\|z_\circ^\perp\|\|\tilde{z}_\circ^\perp\| - \varepsilon^2\bar{C}^2 M_\nu\|z\|\|\tilde{z}_\circ^\perp\| - 2M_\nu\|z_\circ^\perp\|\|z\|\bar{C}\varepsilon. \end{aligned} \tag{3.47}$$

Since by (3.44), $(1 + \varepsilon\bar{C})\|z\| \geq \|z_\circ^\perp\| \geq (1 - \varepsilon\bar{C})\|z\|$ we obtain

$$\langle \mathcal{M}_\nu\, z, \tilde{z}\rangle \geq \left\{\frac{(1 - \varepsilon\bar{C})^2\theta}{2} - \varepsilon\bar{C}M_\nu\big(\varepsilon\bar{C} + 2(1 + \varepsilon\bar{C})\big)\right\}\|z\|\|\tilde{z}\|. \tag{3.48}$$

Therefore, there exists an $\varepsilon_1 = \varepsilon_1(\|\mathcal{C}\|, \lambda^\circ, \theta)$ such that

$$\langle \mathcal{M}_\nu\, z, \tilde{z}\rangle \geq \frac{\theta}{4}\|z\|\|\tilde{z}\|, \quad \forall \varepsilon \leq \varepsilon_1. \tag{3.49}$$

In fact, elementary calculations show that (3.49) is valid for

$$\varepsilon_1 := \frac{1}{8\bar{C}}\min\left\{3, \frac{8\theta}{25\bar{M}}\right\}, \tag{3.50}$$

where $\bar{M}$ is given in (3.42).

Regarding the dependencies of ε_1, we have used (3.45), (3.41), and that $M_\nu = \|\mathcal{M}_\nu\|$ depends on $\|\mathcal{C}\|, \lambda^\circ, \theta$, see (3.42). This concludes the proof. □

We are now prepared to complete the proof of Theorem 3.4:

Proof of Theorem 3.4 Let $\varepsilon \leq \bar{\varepsilon} := \min(\varepsilon_0, \varepsilon_1)$ where ε_0 and ε_1 have been specified in (the proofs of) Lemmata 3.7 and 3.8, see (3.41), (3.50). Hence, the above perturbation results remain uniformly valid for all pairs

$$(u, \nu) \in \mathcal{N} = B(u^\circ, \tau) \times B\left(\lambda^\circ, \frac{\theta}{4\|\mathcal{C}\|}\right) \quad \text{with} \quad \tau = \|u^\circ\|\bar{\varepsilon} = 2\lambda^\circ\bar{\varepsilon}. \tag{3.51}$$

Specifically, we can take in view of (3.41) and (3.50)

$$\tau := 2\lambda^\circ \min\left\{\frac{1}{16\bar{C}}, \frac{\theta}{25\bar{M}\bar{C}}\right\} = \min\left\{\frac{1}{8\|\mathcal{C}\|}, \frac{2\theta}{25(1 + \lambda^\circ\|\mathcal{C}\| + \theta/4)\|\mathcal{C}\|}\right\}. \tag{3.52}$$

We fix now $(\bar{u}, \bar{\nu}) \in \mathcal{N}$. The operator equation: for a given $(g, \zeta) \in \mathbb{U} \times \mathbb{R}$ find the solution $(u, \nu) \in \mathbb{U} \times \mathbb{R}$ of

$$DR(\bar{u}, \bar{\nu})(u, \nu) = (g, \zeta)$$

can be written as the saddle-point problem

$$\begin{aligned} \bar{a}(u,z) + \bar{b}(z,\nu) &= \langle g, z\rangle \quad && \forall z \in \mathbb{U} \\ \bar{b}(u,\alpha) &= \alpha\zeta && \forall \alpha \in \mathbb{R}, \end{aligned} \tag{3.53}$$

where the $(\bar{u}, \bar{\nu})$-dependent bilinear forms read

$$\bar{a}(u,z) := \langle \mathcal{M}_{\bar{\nu}}\, u, z\rangle, \quad \bar{b}(u,\alpha) := -\alpha\langle \mathcal{C}\,\bar{u}, u\rangle. \tag{3.54}$$

Note that we have proceeded in the same spirit as earlier in the proof of Lemma 3.3 (see (3.30)). Here we consider the general case where $(\bar{u}, \bar{\nu}) \in \mathcal{N}$ recovering (3.30) for $(\bar{u}, \bar{\nu}) = (u^\circ, \lambda^\circ)$.

To prove well-posedness of (3.53) we proceed exactly as in Lemma 3.3 and show that (3.53) satisfies the LBB conditions. To that end, it will be convenient to derive bounds for $\|\mathcal{C}\,\bar{u}\|$ first. Since we have chosen $\bar{u} \in B(u^\circ, \tau) = B(u^\circ, \bar{\varepsilon}\|u^\circ\|)$, we can apply Lemma 3.7 for the case $u = \bar{u}$ and $\varepsilon = \bar{\varepsilon} \le \varepsilon_0$ to deduce that

$$\|\mathcal{C}\,\bar{u}\|\|\bar{u}\| \ge \langle \mathcal{C}\,\bar{u}, \bar{u}\rangle \ge \frac{\|\bar{u}\|^2}{2\lambda^\circ}$$

Therefore, by the reverse triangle inequality,

$$\|\mathcal{C}\,\bar{u}\| \ge \frac{\|\bar{u}\|}{2\lambda^\circ} \ge \frac{1}{2\lambda^\circ}(\|u^\circ\| - \|\bar{u} - u^\circ\|) \ge 1 - \bar{\varepsilon}$$

Similarly, by the direct triangle inequality

$$\|\mathcal{C}\,\bar{u}\| \le \|\mathcal{C}\,u^\circ\| + \|\mathcal{C}(\bar{u} - u^\circ)\| \le 2(1 + \bar{\varepsilon}\lambda^\circ\|\mathcal{C}\,\|).$$

To summarize, we have just proven that

$$1 - \bar{\varepsilon} \le \|\mathcal{C}\,\bar{u}\| \le 2(1 + \bar{\varepsilon}\lambda^\circ\|\mathcal{C}\,\|). \tag{3.55}$$

Let us now turn to proving that the LBB conditions are satisfied. We first note that $\bar{a}$ and $\bar{b}$ are bounded bilinear forms since

$$\begin{aligned} \|\bar{a}\| &:= \sup_{u\in\mathbb{U}}\sup_{z\in\mathbb{U}} \frac{\bar{a}(u,z)}{\|u\|\|z\|} \le M_{\bar{\nu}}, \quad \|\bar{b}\| \\ &:= \sup_{\nu\in\mathbb{R}}\sup_{z\in\mathbb{U}} \frac{|b(z,\nu)|}{\|z\||\nu|} \le \|\mathcal{C}\,\bar{u}\| \le 2(1 + \bar{\varepsilon}\lambda^\circ\|\mathcal{C}\,\|). \end{aligned} \tag{3.56}$$

Next, denoting

$$Z = \{w \in \mathbb{U} \,:\, \bar{b}(w,\alpha) = 0,\ \forall \alpha \in \mathbb{R}\} = \{w \in \mathbb{U} \,:\, \langle \mathcal{C}\,\bar{u}, w\rangle = 0\} = \mathrm{span}\{\mathcal{C}\,\bar{u}\}^\perp,$$

applying Lemma 3.8 to $\mathcal{M}_{\bar{\nu}}$ yields inf-sup stability of $\mathcal{M}_{\bar{\nu}}$ on Z,

$$\left\langle \mathcal{M}_{\bar{\nu}}\, w, w' \right\rangle \geq \frac{\theta}{4} \|w\| \|w'\|, \quad \forall w, w' \in \operatorname{span}\{\mathcal{C}\,\bar{u}\}^{\perp}.$$

In addition, using the lower bound from (3.55),

$$\inf_{\alpha \in \mathbb{R}} \sup_{z \in \mathbb{U}} \frac{|\bar{b}(z,\alpha)|}{\|z\||\alpha|} = \sup_{z \in \mathbb{U}} \frac{\langle \mathcal{C}\,\bar{u}, z\rangle}{\|z\|} = \|\,\mathcal{C}\,\bar{u}\| \geq 1 - \bar{\varepsilon}.$$

This proves that the LBB conditions are satisfied. Thus, problem (3.53) is well-posed, and invoking standard stability estimates e.g. from [10], we have

$$\begin{aligned}
\|u\| &\leq \frac{4}{\theta}\|g\| + \frac{1}{1-\bar{\varepsilon}}\left(1 + 4\frac{M_{\bar{\nu}}}{\theta}\right)|\xi| \\
&\leq \underbrace{\left(\frac{16}{\theta^2} + \frac{1}{(1-\bar{\varepsilon})^2}\left(1 + 4\frac{M_{\bar{\nu}}}{\theta}\right)^2\right)^{1/2}}_{:=\beta_1} \|(g,\xi)\|, \\
|\nu| &\leq \frac{1}{1-\bar{\varepsilon}}\left(1 + 4\frac{M_{\bar{\nu}}}{\theta}\right)\|g\| + \frac{M_{\bar{\nu}}}{(1-\bar{\varepsilon})^2}\left(1 + 4\frac{M_{\bar{\nu}}}{\theta}\right)|\nu| \\
&\leq \left(\frac{1}{(1-\bar{\varepsilon})^2}\left(1 + 4\frac{M_{\bar{\nu}}}{\theta}\right)^2 + \frac{M_{\bar{\nu}}^2}{(1-\bar{\varepsilon})^4}\left(1 + 4\frac{M_{\bar{\nu}}}{\theta}\right)^2\right)^{1/2} \|(g,\xi)\| \\
&= \underbrace{\frac{1}{(1-\bar{\varepsilon})}\left(1 + 4\frac{M_{\bar{\nu}}}{\theta}\right)\left(1 + \frac{M_{\bar{\nu}}}{(1-\bar{\varepsilon})^2}\right)^{1/2}}_{:=\beta_2} \|(g,\xi)\|.
\end{aligned}$$

Thus, we conclude that

$$\|DR(\bar{u},\bar{\nu})^{-1}\|_{\mathcal{L}(\tilde{\mathbb{U}},\tilde{\mathbb{U}})} \leq \bar{\beta} = \sqrt{\beta_1^2 + \beta_2^2}, \quad \forall (\bar{u},\bar{\nu}) \in \mathcal{N},$$

where (3.36) follows from substituting $\varepsilon = \frac{\tau}{2\lambda^\circ}$. Thanks to the above inequalities, and recalling from (3.46) that

$$M_{\bar{\nu}} \leq \bar{M} := 1 + \lambda^\circ \|\,\mathcal{C}\,\| + \frac{\theta}{4}, \quad \forall \bar{\nu} \in B\left(\lambda^\circ, \frac{\theta}{4\|\,\mathcal{C}\,\|}\right),$$

the assertion follows from (3.52).

3.5 Convergence of the Newton Scheme

In this section, we prove that the Newton scheme (3.3) is locally quadratically convergent to (u°, λ°). Proving this will require leveraging Lipschitz continuity of DR, a property which we record in the next lemma.

Lemma 3.9 *The derivative DR is Lipschitz continuous, and*

$$\begin{aligned}\|DR(u_1, \nu_1) - DR(u_2, \nu_2)\|_{\mathcal{L}(\mathbb{U}\times\mathbb{R}, \mathbb{U}\times\mathbb{R})} \le \gamma \|(u_1, \nu_1) \\ -(u_2, \nu_2)\|_{\tilde{\mathbb{U}}}, \quad \forall\ (u_1, \nu_1), (u_2, \nu_2) \in \mathbb{U}\times\mathbb{R},\end{aligned} \tag{3.57}$$

where the Lipschitz constant γ can be bounded by $\gamma \le \sqrt{2}\| \mathcal{C} \|$.

Proof Since $DR(\cdot, \cdot)$ is linear in both arguments, it is Lipschitz continuous if and only if it is bounded. A repeated combination of the triangle inequality and the Cauchy–Schwarz inequality yields the bound. □

We are now in position to prove convergence of the Newton scheme. The main arguments in the development are actually classical, and can be found in different references (see, e.g., [21, 42]). It will be convenient to denote the ball in $\tilde{\mathbb{U}}$ centered at (u°, λ°) of radius ω as

$$K_\omega(u^\circ, \lambda^\circ) := B((u^\circ, \lambda^\circ), \omega) = \{(u, \mu) \in \tilde{\mathbb{U}} \ :\ \|(u^\circ, \lambda^\circ) - (u, \mu)\| < \omega\}.$$

Theorem 3.10 *(Convergence of the Newton scheme) Let $\mathcal{N}$ be the neighborhood defined in Theorem 3.4. Let $\omega > 0$ be sufficiently small such that*

$$K_\omega(u^\circ, \lambda^\circ) \subset \mathcal{N}, \quad \text{and} \quad \omega \le \frac{2}{\bar{\beta}\gamma}, \tag{3.58}$$

where $\bar{\beta}$ is defined in Theorem (3.4), and γ is the Lipschitz constant from Lemma 3.9.

If the initial guess $(u_0, \lambda_0) \in K_\omega(u^\circ, \lambda^\circ)$, then the sequence of Newton iterates $((u_n, \lambda_n))_{n=0}^\infty$ stays in $K_\omega(u^\circ, \lambda^\circ)$ and has quadratic convergence,

$$\|(u_{n+1}, \lambda_{n+1}) - (u^\circ, \lambda^\circ)\| \le \frac{\bar{\beta}\gamma}{2}\|(u_n, \lambda_n) - (u^\circ, \lambda^\circ)\|^2. \tag{3.59}$$

Remark 3.11 Notice that, in view of Remark 3.5, the condition on ω in (3.58) agrees in essence with the condition on the neighborhood $\mathcal{N}$ and therefore introduces no severe additional constraints.

Proof We sketch it for the convenience of the reader since the proof consists of "lifting" standard arguments to an operator level. By induction, for a given $n \ge 0$, we assume that $(u_n, \lambda_n) \in K_\omega(u^\circ, \lambda^\circ)$ (the case $n = 0$ being true by assumption). We show that there is a contraction in the error at the next step, and that $(u_{n+1}, \lambda_{n+1}) \in K_\omega(u^\circ, \lambda^\circ)$.

To that end, since $R(u^\circ, \lambda^\circ) = 0$, we have

$$\begin{aligned}
&(u_{n+1}, \lambda_{n+1}) - (u^\circ, \lambda^\circ) \\
&\quad = (u_n, \lambda_n) - (u^\circ, \lambda^\circ) - DR(u_n, \lambda_n)^{-1} R(u_n, \lambda_n) \\
&\quad = (u_n, \lambda_n) - (u^\circ, \lambda^\circ) - DR(u_n, \lambda_n)^{-1} \left(R(u_n, \lambda_n) - R(u^\circ, \lambda^\circ)\right) \\
&\quad = DR(u_n, \lambda_n)^{-1} \left(R(u^\circ, \lambda^\circ) - R(u_n, \lambda_n)\right. \\
&\qquad \left. - DR(u_n, \lambda_n)((u^\circ, \lambda^\circ) - (u_n, \lambda_n))\right).
\end{aligned}$$

Note that the derivative $DR(u_n, \lambda_n)^{-1}$ is well defined by Theorem 3.4 since $(u_n, \lambda_n) \in K_\omega(u^\circ, \lambda^\circ) \subset \mathcal{N}$. Taking norms gives

$$\begin{aligned}
\|(u_{n+1}, \lambda_{n+1}) - (u^\circ, \lambda^\circ)\| \le \bar{\beta} \|&R(u^\circ, \lambda^\circ) - R(u_n, \lambda_n) \\
&- DR(u_n, \lambda_n)((u^\circ, \lambda^\circ) - (u_n, \lambda_n))\|. \qquad (3.60)
\end{aligned}$$

To derive a bound for the right-hand side of (3.60), we introduce the segment

$$\ell_n(t) := (u_n, \lambda_n) + t((u^\circ, \lambda^\circ) - (u_n, \lambda_n)), \quad \forall t \in [0, 1]$$

joining (u_n, λ_n) and (u°, λ°), and we consider

$$\phi(t) := R \circ \ell_n(t)$$

which is the restriction of R to $\ell_n(t)$. Note that since

$$\phi'(t) = DR\left((u_n, \lambda_n) + t((u^\circ, \lambda^\circ) - (u_n, \lambda_n))\right)\left((u^\circ, \lambda^\circ) - (u_n, \lambda_n)\right),$$

we have

$$\begin{aligned}
&\|R(u^\circ, \lambda^\circ) - R(u_n, \lambda_n) - DR(u_n, \lambda_n)((u^\circ, \lambda^\circ) - (u_n, \lambda_n))\| \\
&\quad = \|\phi(1) - \phi(0) - \phi'(0)\| \\
&\quad \le \int_0^1 \|\phi'(t) - \phi'(0)\| \mathrm{d}t \le \frac{\gamma}{2} \|(u^\circ, \lambda^\circ) - (u_n, \lambda_n)\|^2.
\end{aligned}$$

Inserting this last inequality into (3.60) yields (3.59).

We finally prove that $(u_{n+1}, \lambda_{n+1}) \in K_\omega((u^\circ, \lambda^\circ))$. This follows from the inequalities

$$\begin{aligned}
\|(u_{n+1}, \lambda_{n+1}) - (u^\circ, \lambda^\circ)\| &\le \frac{\bar{\beta}\gamma}{2} \|(u^\circ, \lambda^\circ) - (u_n, \lambda_n)\| \|(u^\circ, \lambda^\circ) - (u_n, \lambda_n)\| \\
&\le \frac{\bar{\beta}\gamma\omega}{2} \|(u^\circ, \lambda^\circ) - (u_n, \lambda_n)\| \\
&\le \|(u^\circ, \lambda^\circ) - (u_n, \lambda_n)\| < \omega.
\end{aligned}$$

□

4 Perturbed Newton Scheme

4.1 Goals and Notation

The idealized Newton scheme (3.3) is formulated on a continuous level (see (CC1) in Sect. 1.3), i.e., the Newton updates (3.4) require exact inversions of the system (3.4). According to (CC2), (CC3) in Sect. 1.3, we aim at a practical realization based on approximating these updates within judicious controllable tolerances. Specifically, for any $(\bar{u}, \bar{v}) \in \tilde{\mathbb{U}}$, any target accuracy $\eta > 0$, and any right-hand side $f \in \mathbb{W}' \times \mathbb{R}$, numerical solver has to deliver an approximation $(w_\eta, \alpha_\eta) \in \tilde{\mathbb{U}}$ to the operator equation

$$DR(\bar{u}, \bar{v})(w, \alpha) = f.$$

We denote a numerical solver that realizes accuracy η with output (w_η, α_η) by

$$[DR(\bar{u}, \bar{v})^{-1}, f; \eta] \to (w_\eta, \alpha_\eta), \tag{4.1}$$

meaning that

$$\|(w_\eta, \alpha_\eta) - DR(\bar{u}, \bar{v})^{-1} f\| \le \eta.$$

We postpone discussing possible realizations of $[DR(\bar{u}, \bar{v})^{-1}, f; \eta]$. Assuming for the moment to have such a scheme at hand, we discuss in Sect. 4.2 first for which (dynamically adjusted) tolerances $\eta = \eta_n$ a corresponding inexact Newton method converges at a linear or even quadratic rate, see (CC2), Sect. 1.3.

Finally, we discuss in Sect. 4.4 strategies for numerically realizing such accuracy controlled inexact Newton updates.

4.2 Scheme and Convergence Analysis

Recall from (3.3) and (3.4) that the exact Newton iteration produces a sequence $(u_n, \lambda_n)_{n\ge 0}$ defined by

$$u_{n+1} = u_n + \delta_n^{(u)}, \quad \lambda_{n+1} = \lambda_n + \delta_n^{(\lambda)}, \quad n \in \mathbb{N}_0, \tag{4.2}$$

with

$$(\delta_n^{(u)}, \delta_n^{(\lambda)}) = -DR(u_n, \lambda_n)^{-1} R(u_n, \lambda_n). \tag{4.3}$$

The inexact iteration gives rise to a sequence $(\bar{u}_n, \bar{\lambda}_n)_{n\ge 0}$ defined by

$$\bar{u}_{n+1} = \bar{u}_n + \bar{h}_n^{(u)}, \quad \bar{\lambda}_{n+1} = \bar{\lambda}_n + \bar{h}_n^{(\lambda)}, \quad n \in \mathbb{N}_0, \tag{4.4}$$

where

$$(\bar{h}_n^{(u)}, \bar{h}_n^{(\lambda)}) = [DR(\bar{u}_n, \bar{\lambda}_n)^{-1}, -R(\bar{u}_n, \bar{\lambda}_n); \eta_n] \tag{4.5}$$

for some accuracy tolerance η_n. The computable updates $(\bar{h}_n^{(u)}, \bar{h}_n^{(\lambda)})$ approximate the solution to the operator equation

$$DR(\bar{u}_n, \bar{\lambda}_n)(h_n^{(u)}, h_n^{(\lambda)}) = -R(\bar{u}_n, \bar{\lambda}_n).$$

By definition of $[DR(\bar{u}_n, \bar{\lambda}_n)^{-1}, -R(\bar{u}_n, \bar{\lambda}_n); \eta_n]$, we thus have

$$\|(\bar{h}_n^{(u)}, \bar{h}_n^{(\lambda)}) - (h_n^{(u)}, h_n^{(\lambda)})\| \leq \eta_n. \tag{4.6}$$

The next theorem gives an upper bound for the value of η_n that preserves the same quadratic convergence rate as in the exact Newton scheme. For its proof, it will be convenient to introduce notation for the convergence error of the exact and perturbed schemes

$$e_n := \|(u_n, \lambda_n) - (u^\circ, \lambda^\circ)\|, \qquad \bar{e}_n := \|(\bar{u}_n, \bar{\lambda}_n) - (u^\circ, \lambda^\circ)\|. \tag{4.7}$$

Theorem 4.1 *For the neighborhood* $\mathcal{N} = B(u^\circ, \tau) \times B\left(\lambda^\circ, \frac{\theta}{4\|\mathcal{C}\|}\right)$, $\tau = \|u^\circ\|\bar{\varepsilon}$, *defined in* (3.51), *let* $\omega > 0$ *be sufficiently small such that*

$$K_\omega(u^\circ, \lambda^\circ) \subset \mathcal{N}, \quad \textit{and} \quad \omega \leq \frac{1}{3\bar{\beta}\gamma}. \tag{4.8}$$

Then, if $(\bar{u}_0, \bar{\lambda}_0) \in K_\omega(u^\circ, \lambda^\circ)$,

$$\bar{e}_{n+1} \leq \eta_n + \frac{\bar{\beta}\gamma}{2}(\bar{e}_n^2 + 2e_n^2). \tag{4.9}$$

In addition, if

$$\eta_n \leq \min\left(\frac{\omega}{2}, \frac{\bar{\beta}\gamma}{2}\bar{e}_n^2\right), \tag{4.10}$$

all iterates remain in $K_\omega(u^\circ, \lambda^\circ)$, *i.e.,*

$$(u_n, \lambda_n) \in K_\omega(u^\circ, \lambda^\circ), \quad \forall n \geq 0. \tag{4.11}$$

Finally, we retain quadratic convergence as in the unperturbed case, namely as soon as n_0 *satisfies*

$$\bar{\beta}\gamma\bar{e}_{n_0}^2 \leq \omega, \tag{4.12}$$

one has

$$\bar{e}_{n+1} \leq 4\bar{\beta}\gamma\bar{e}_n^2, \quad n \geq n_0. \tag{4.13}$$

Proof For any $n \geq 0$, we have

$$\begin{aligned}\bar{e}_{n+1} = \|(\bar{u}_{n+1}, \bar{\lambda}_{n+1}) - (u^\circ, \lambda^\circ)\| &\leq \|(\bar{u}_{n+1}, \bar{\lambda}_{n+1}) - (u_{n+1}, \lambda_{n+1})\| \\ &\quad + \|(u_{n+1}, \lambda_{n+1}) - (u^\circ, \lambda^\circ)\| \qquad (4.14)\\ &\leq \|(\bar{u}_{n+1}, \bar{\lambda}_{n+1}) - (u_{n+1}, \lambda_{n+1})\| + \frac{\bar{\beta}\gamma}{2}e_n^2, \qquad (4.15)\end{aligned}$$

where we have used (3.59) in the last inequality. We next build a bound for the first term of the inequality. We have

$$\begin{aligned}&\|(\bar{u}_{n+1}, \bar{\lambda}_{n+1}) - (u_{n+1}, \lambda_{n+1})\| \\ &\quad = \|(\bar{u}_n, \bar{\lambda}_n) + (\bar{h}_n^{(u)}, \bar{h}_n^{(\lambda)}) - (u_n, \lambda_n) - (\delta_n^{(u)}, \delta_n^{(\lambda)})\| \\ &\quad \leq \|(\bar{h}_n^{(u)}, \bar{h}_n^{(\lambda)}) - (h_n^{(u)}, h_n^{(\lambda)})\| + \|(\bar{u}_n, \bar{\lambda}_n) - DR(\bar{u}_n, \bar{\lambda}_n)^{-1}R(\bar{u}_n, \bar{\lambda}_n) - (u_n, \lambda_n) \\ &\qquad + DR(u_n, \lambda_n)^{-1}R(u_n, \lambda_n)\| \\ &\quad \leq \eta_n + \|(\bar{u}_n, \bar{\lambda}_n) - (u^\circ, \lambda^\circ) - DR(\bar{u}_n, \bar{\lambda}_n)^{-1}\left(R(\bar{u}_n, \bar{\lambda}_n) - R(u^\circ, \lambda^\circ)\right)\| \\ &\qquad + \|(u_n, \lambda_n) - (u^\circ, \lambda^\circ) - DR(u_n, \lambda_n)^{-1}\left(R(u_n, \lambda_n) - R(u^\circ, \lambda^\circ)\right)\|,\end{aligned}$$

where we have added and subtracted (u°, λ°) and used that $R(u^\circ, \lambda^\circ) = 0$ to derive the last inequality. Now, by the same arguments as in the proof of Theorem 3.10, we conclude that

$$\begin{aligned}&\|(\bar{u}_n, \bar{\lambda}_n) - (u^\circ, \lambda^\circ) - DR(\bar{u}_n, \bar{\lambda}_n)^{-1}\left(R(\bar{u}_n, \bar{\lambda}_n) - R(u^\circ, \lambda^\circ)\right)\| \leq \frac{\bar{\beta}\gamma}{2}\|(\bar{u}_n, \bar{\lambda}_n) \\ &\quad -(u^\circ, \lambda^\circ)\|^2 = \frac{\bar{\beta}\gamma}{2}\bar{e}_n^2,\end{aligned}$$

and

$$\|(u_n, \lambda_n) - (u^\circ, \lambda^\circ) - DR(u_n, \lambda_n)^{-1}\left(R(u_n, \lambda_n) - R(u^\circ, \lambda^\circ)\right)\| \leq \frac{\bar{\beta}\gamma}{2}e_n^2.$$

Hence, we obtain

$$\|(\bar{u}_{n+1}, \bar{\lambda}_{n+1}) - (u_{n+1}, \lambda_{n+1})\| \leq \eta_n + \frac{\bar{\beta}\gamma}{2}(\bar{e}_n^2 + e_n^2). \tag{4.16}$$

Inserting this inequality into (4.15) yields (4.9). Since by assumption, $(\bar{u}_0, \bar{\lambda}_0) = (u_0, \lambda_0) \in K_\omega((u^\circ, \lambda^\circ))$, we infer inductively from (4.8) and (4.10) that

$$\bar{e}_{n+1} \leq \eta_n + \frac{\bar{\beta}\gamma}{2}\big(\bar{e}_n^2 + 2e_n^2\big) = \eta_n + \frac{3\bar{\beta}\gamma\omega}{2}\omega \leq \eta_n + \frac{\omega}{2} < \omega, \tag{4.17}$$

which is (4.11).

Concerning inequality (4.13), substituting (4.10) into (4.16), yields under the provision (4.12) the asserted quadratic convergence (4.13). □

Realizing an update accuracy of the order $\eta_n \sim \bar{e}_n^2$ may actually entail a bit of a challenge for a numerical realization. It is important to note though that perhaps of equal practical importance is the fact that a less demanding tolerance η_n would still give rise to a robust first-order scheme with control on the quantitative error reduction rate as detailed next.

Remark 4.2 Suppose that for any given $0 < \zeta \ll 1$, n_1 is large enough to ensure that $3\bar{\beta}\gamma\bar{e}_n \leq \zeta/2$, $n \geq n_1$, then whenever

$$\eta_n = \min\left\{\frac{\omega}{2}, \frac{\zeta}{2}\bar{e}_n\right\}, \tag{4.18}$$

(4.9) says that for $n \geq n_1$ $\bar{e}_{n+1} \leq \left(\frac{\zeta}{2} + \frac{3\bar{\beta}\gamma}{2}\bar{e}_n\right)\bar{e}_n$, which shows that

$$\bar{e}_{n+1} \leq \zeta\bar{e}_n. \tag{4.19}$$

Hence an update tolerance proportional to the current accuracy, with sufficiently small proportionality factor, gives rise to a linear error reduction with a reduction factor as small as one wishes.

The significance of these observations lies in the following familiar consequence. Suppose one has $\bar{e}_{n+1} \leq c\bar{e}_n^p$ for some $0 < c < \infty$ and $p \geq 1$. Then, triangle inequality yields

$$(1 - c\bar{e}_n^{p-1})\bar{e}_n \leq \|\bar{u}_{n+1} - \bar{u}_n\| \leq (1 + c\bar{e}_n^{p-1})\bar{e}_n. \tag{4.20}$$

Thus, for quadratic convergence $p = 2$ and any fixed $c < \infty$, and for linear convergence $p = 1$ but $c \leq 1/2$ say, one has (for n sufficiently large)

$$\frac{1}{2}\bar{e}_n \leq \|\bar{u}_{n+1} - \bar{u}_n\| \leq \frac{3}{2}\bar{e}_n. \tag{4.21}$$

which means that the computable a posteriori quantity $\|\bar{u}_{n+1} - \bar{u}_n\|$ provides a tight error bound for the current approximation $\bar{u}_n$.

4.3 A Posteriori Estimation of $\bar{e}_n$

The remarks at the end of the previous section show that a current error $\bar{e}_n$ can be assessed through a computable a posteriori quantity provided that either (4.13) or (4.19) (for sufficiently small ζ) hold. This in turn, hinges on two pillars. First, one needs to estimate $\bar{e}_n$ in order to determine a suitable η_n (we cannot use (4.20) at this stage because $\bar{u}_{n+1}$ has yet to be computed). Second, once $\bar{e}_n$ and hence η_n is known, one needs to solve the Newton update problem (4.5) with the tolerance η_n. We defer this latter issue to the next section and discuss here the first one through deriving an a posteriori error bound that does not require the subsequent approximation. The rationale is, in principle, simple. We have already shown that the linearization of $R(u, \nu) = 0$ is well-posed in a certain neighborhood of (u°, λ°) which can be used as follows.

Proposition 4.3 *There exists a constant $0 < \overline{C} < \infty$, depending only on the problem parameters $\| \mathcal{C} \|$, $\bar{M}$ such that*

$$\begin{aligned} \bar{\beta} \| (u, \nu) - (u^\circ, \lambda^\circ) \| &\le \| R(u, \nu) \| \\ &\le \overline{C} \| (u, \nu) - (u^\circ, \lambda^\circ) \|, \quad \forall\ (u, \nu) \in K_\omega(u^\circ, \lambda^\circ). \end{aligned} \tag{4.22}$$

Proof Arguing as before, for any $(u, \nu) \in K_\omega((u^\circ, \lambda^\circ))$ we introduce the segment

$$\ell(t) = (u^\circ, \lambda^\circ) + t((u, \nu) - (u^\circ, \lambda^\circ)), \quad \forall t \in [0, 1]$$

and the restriction of R along that segment

$$\phi(t) = R \circ \ell(t), \quad \forall t \in [0, 1].$$

It follows that

$$\begin{aligned} \| R(u, \nu) \| = \| R(u, \nu) - R(u^\circ, \lambda^\circ) \| = \| \phi(1) - \phi(0) \| = \| \int_0^1 \phi'(t) \mathrm{d}t \| \\ \le \max_{t \in [0,1]} |\phi'(t)| \le \overline{C} \| (u, \nu) - (u^\circ, \lambda^\circ) \|, \end{aligned} \tag{4.23}$$

where

$$\overline{C} := \sup_{t \in [0,1]} \| DR((u^\circ, \lambda^\circ) + t((u, \nu) - (u^\circ, \lambda^\circ)) \|$$

depends on $\bar{M}$, $\| \mathcal{C} \|$.

To obtain also a lower bound, we apply the mean value value theorem to ϕ over $[0, 1]$ to deduce that there exists $\bar{t} \in (0, 1)$ such that

$$R(u, \nu) - R(u^\circ, \lambda^\circ) = \phi(1) - \phi(0) = \phi'(\bar{t}) = DR(\ell(\bar{t}))((u, \nu) - (u^\circ, \lambda^\circ)).$$

Since $\ell(\bar{t}) \in K_\omega((u^\circ, \lambda^\circ))$, Theorem 3.4 guarantees injectivity of $DR(\ell(\bar{t}))$, from which it follows that

$$(u, \nu) - (u^\circ, \lambda^\circ) = DR^{-1}(\ell(\bar{t}))R(u, \nu).$$

Taking norms and using (3.34) yields the lower bound

$$\|(u, \nu) - (u^\circ, \lambda^\circ) \leq \bar{\beta}^{-1}\|R(u, \nu)\|.$$

As a side remark, note that one can actually derive that the exact value of $\bar{t}$ is $1/2$ by exactly integrating $\phi(1) - \phi(0) = \int_0^1 \phi'(t)\mathrm{d}t$ with the expression for $\phi'(t)$ which reads

$$\begin{aligned}\phi'(t) &= DR(\ell(t))((u, \nu) - (u^\circ, \lambda^\circ)) \\ &= \begin{pmatrix} (I - (\lambda^\circ + t(\nu - \lambda^\circ))\,\mathcal{C}(u - u^\circ) - \nu\,\mathcal{C}(u^\circ + t(u - u^\circ)) \\ -\langle \mathcal{C}(u^\circ + t(u - u^\circ)), u - u^\circ)\rangle \end{pmatrix}.\end{aligned}$$

□

Proposition 4.3 implies that the accuracy provided by each perturbed step can be assessed through $\|R(u_n, \lambda_n)\|$, and this quantity can be used to steer the target accuracy of subsequent calculations. Hence, at this stage, things boil down to computing accuracy controlled Newton updates. We sketch the essence of a corresponding scheme in the next section, leaving details to forthcoming work with numerical realizations.

4.4 Towards a Numerical Realization

Recall that in Sect. 4.2 we have identified appropriate accuracy tolerances and in Sect. 4.3 we have shown how to assess the accuracy of a given approximation. The objective of this section is to provide the analytical basis for numerically realizing approximate Newton updates within a given error tolerance. The basic strategy is to reduce this task to judicious applications of a routine

$$(w, \eta) \in \mathbb{U} \times \mathbb{R}_+ \mapsto [\mathcal{C}, w; \eta] \quad \text{such that} \quad \big\| \mathcal{C}\, w - [\mathcal{C}, w; \eta]\big\| \leq \eta. \tag{4.24}$$

This routine, in turn, can be realized with the aid of an error controlled application scheme $[\mathcal{F}, z; \eta]$, (i.e., $\|\mathcal{F}z - [\mathcal{F}, z; \eta]\| \leq \eta$) in combination with an error controlled solver for the source problem $[\mathcal{B}^{-1}, q; \eta]$, providing an η-accurate approximation to the solution of $\mathcal{B}u = q$, i.e., $\|u - [\mathcal{B}^{-1}, q; \eta]\| \leq \eta$. Routines of the form $[\mathcal{F}, z; \eta]$ have been discussed in [15] for non-trivial kernels, using wavelet compression or low-rank approximation. The routine $[\mathcal{B}^{-1}, q; \eta]$ is the overall main result of [15]. So it is justified for the purpose of the subsequent discussion to claim availability of these two routines.

Then, the following can be easily verified and will be used repeatedly.

Remark 4.4 With the routines $[\mathcal{F}, \cdot; \cdot]$ and $[\mathcal{B}^{-1}, \cdot; \cdot]$ at hand

$$[\mathcal{C}, f; \eta] := [\mathcal{B}^{-1}, [\mathcal{F}, f; \eta_1]; \eta_2] \tag{4.25}$$

satisfies $\|[\mathcal{C}, f; \eta] - \mathcal{C} f\| \leq \eta$ provided that $\|\mathcal{B}^{-1}\|\eta_1 + \eta_2 \leq \eta$.

There are actually several pathways of employing $[\mathcal{C}, f; \eta]$. The subsequent discussion is to shed light on intrinsic obstructions that arise along the way as well as to line out remedies.

To this end, recall from (3.5), (3.7) that, given the exact Newton iterates (u_n, λ_n), the exact Newton updates can be obtained via block elimination as follows

$$\begin{aligned}(\delta_n^{(u)}, \delta_n^{(\lambda)}) &= -DR(u_n, \lambda_n)^{-1} R(u_n, \lambda_n) \\ &\iff \begin{cases} \delta_n^{(u)} &= \delta_n^{(\lambda)} \mathcal{M}_{\lambda_n}^{-1} \mathcal{C} u_n - u_n \\ \delta_n^{(\lambda)} &= \dfrac{1 + \langle \mathcal{C} u_n, u_n \rangle - \|\mathcal{C} u_n\|^2/2}{\left\langle \mathcal{C} u_n, \mathcal{M}_{\lambda_n}^{-1} \mathcal{C} u_n \right\rangle}. \end{cases}\end{aligned} \tag{4.26}$$

A first natural approach is to realize the scheme (4.5) by approximately computing the quantities in the right-hand part of (4.26). However, in addition to $[\mathcal{C}, f; \eta]$ this would require a routine: for every $f \in \mathbb{U}$, $\nu \in \mathbb{R}$ and target tolerance $\eta > 0$, compute approximations to $\mathcal{M}_{\lambda_m}^{-1} \mathcal{C} u_n$. Since we have an accuracy controlled of $\mathcal{C}$ at hand, it would be natural to develop a routine

$$w_\eta = [\mathcal{M}_\nu^{-1}, f; \eta] \quad \text{such that} \quad \|w_\eta - \mathcal{M}_\nu^{-1} f\| \leq \eta. \tag{4.27}$$

The obvious problem is that the operator $\mathcal{M}_\nu$ becomes increasingly harder to invert when ν approaches λ°, as the values λ_n would do. We postpone a way of circumventing these difficulties, provided that (a good lower bound for) the spectral gap Δ (see (3.24)) is known, to the end of this section.

We address therefore first another option that takes advantage of Theorem 3.4. In fact, well-posedness of the saddle point problem just means that $DR(\bar{u}, \bar{\nu})$ is a $\mathbb{U}$-isomorphism in a neighborhood of (u°, λ°), depending on λ°, $\|\mathcal{C}\|$ and θ. Hence, errors are equivalent to residuals in the sense that

$$\begin{aligned}\|(\bar{w}, \bar{\nu}) - (\delta_n^{(u^\circ)}, \delta_n^{(\lambda^\circ)})\|^2 &\eqsim \|DR(u_n, \lambda_n)(\bar{w}, \bar{\nu}) + R(u_n, \lambda_n)\|^2 \\ &= \|\mathcal{M}_{\lambda_n} \bar{w} - \bar{\nu} \mathcal{C} u_n + \mathcal{M}_{\lambda_n} u_n\|^2 \\ &\quad + \Big(1 - \frac{\|\mathcal{C} u_n\|^2}{2} - \langle \mathcal{C} u_n, \bar{w} \rangle\Big)^2.\end{aligned} \tag{4.28}$$

Hence, we can view the update $(\delta_n^{(u^\circ)}, \delta_n^{(\lambda^\circ)})$ as the unique minimizer of a quadratic functional. Note that, just using the routine (4.25), we can compute the quantities

$$f_n := \mathcal{C} u_n, \quad g_n := \mathcal{M}_{\lambda_n} u_n, \quad s_n := 1 - \frac{\|f_n\|^2}{2}$$

within any desirable accuracy which leaves us with solving

$$\begin{aligned}(\bar\delta_n^{(u)}, \bar\delta_n^{(\lambda)}) &= \operatorname{argmin}_{\bar w,\bar\nu} Q(\bar w, \bar\nu),\\ Q(\bar w, \bar\nu) &:= \frac{1}{2}\Big\{\|\mathcal{M}_{\lambda_n}\bar w - \bar\nu f_n + g_n\|^2 + \big(s_n - \langle f_n, \bar w\rangle\big)^2\Big\}.\end{aligned} \tag{4.29}$$

Note that the exact update component $\delta_n^{(u)}$ belongs to $\langle \mathcal{C}\, u_n\rangle^\perp$. We know from Lemma 3.8 that $\mathcal{M}_{\lambda_n}$ is well-conditioned on $\langle \mathcal{C}\, u_n\rangle^\perp$. Hence, a gradient descent in $\langle \mathcal{C}\, u_n\rangle^\perp$ converges rapidly towards the minimizer. So, we do not suggest to minimize Q over a fixed finite-dimensional trial space but seek to realize an accuracy controlled gradient descent in function space.

To that end, since $2Q(\bar w, \bar\nu) = \|\mathcal{M}_{\lambda_n}\bar w - \bar\nu f_n\|^2 + 2\big\langle \mathcal{M}_{\lambda_n}\bar w - \bar\nu f_n, g_n\big\rangle + \|g_n\|^2 + s_n^2 + \langle f_n, \bar w\rangle^2 - 2s_n\langle f_n, \bar w\rangle$, it suffices to minimize

$$P(\bar w, \bar\nu) := \frac{1}{2}\|\mathcal{M}_{\lambda_n}\bar w - \bar\nu f_n\|^2 + \big\langle \mathcal{M}_{\lambda_n}\bar w - \bar\nu f_n, g_n\big\rangle + \frac{1}{2}\langle f_n, \bar w\rangle^2 - s_n\langle f_n, \bar w\rangle.$$

Although $\nabla P(\bar w, \bar\nu)$ can be determined along the same lines, as stated earlier for DR, we repeat the specific calculation for the convenience of the reader. Straightforward manipulations yield

$$\begin{aligned}&2\big(P(\bar w + th, \bar\nu + t\omega) - P(\bar w, \bar\nu)\big)\\ &\quad= t^2\big\langle \mathcal{M}_{\lambda_n} h - \omega f_n, \mathcal{M}_{\lambda_n} h - \omega f_n\big\rangle + 2t\big\langle \mathcal{M}_{\lambda_n} h - \omega f_n, \mathcal{M}_{\lambda_n}\bar w - \bar\nu f_n\big\rangle\\ &\qquad+ 2t\big\langle \mathcal{M}_{\lambda_n} h - \omega f_n, g_n\big\rangle + t^2\langle f_n, h\rangle^2 + 2t\langle f_n, \bar w\rangle\langle f_n, h\rangle - 2s_n t\langle f_n, h\rangle.\end{aligned}$$

Thus

$$\begin{aligned}&\lim_{t\to 0}\frac{1}{t}\Big\{P(\bar w + th, \bar\nu + t\omega) - P(\bar w, \bar\nu)\Big\}\\ &\quad= \big\langle \mathcal{M}_{\lambda_n} h - \omega f_n, \mathcal{M}_{\lambda_n}\bar w - \bar\nu f_n + g_n\big\rangle + (\langle f_n, \bar w\rangle - s_n)\langle f_n, h\rangle\\ &\quad= \big\langle \mathcal{M}_{\lambda_n}^*(\mathcal{M}_{\lambda_n}\bar w - \bar\nu f_n + g_n) + (\langle f_n, \bar w\rangle - s_n) f_n, h\big\rangle\\ &\qquad- \omega\big\langle f_n, \mathcal{M}_{\lambda_n}\bar w - \bar\nu f_n + g_n\big\rangle.\end{aligned}$$

Hence,

$$\big\langle \nabla P(\bar w, \bar\nu), (h, \omega)\big\rangle = \left\langle \begin{pmatrix} \mathcal{M}_{\lambda_n}^*(\mathcal{M}_{\lambda_n}\bar w - \bar\nu f_n + g_n) + (\langle f_n, \bar w\rangle - s_n) f_n\\ -\big\langle f_n, \mathcal{M}_{\lambda_n}\bar w - \bar\nu f_n + g_n\big\rangle\end{pmatrix}, \begin{pmatrix} h\\ \omega\end{pmatrix}\right\rangle,$$

which says that

$$D(\bar w, \bar\nu) := -\nabla P(\bar w, \bar\nu) = \begin{pmatrix} -\mathcal{M}_{\lambda_n}^*(\mathcal{M}_{\lambda_n}\bar w - \bar\nu f_n + g_n) + (\langle f_n, \bar w\rangle - s_n) f_n\\ \big\langle f_n, \mathcal{M}_{\lambda_n}\bar w - \bar\nu f_n + g_n\big\rangle\end{pmatrix} \tag{4.30}$$

is the direction of steepest descent in $\tilde{\mathbb{U}} = \mathbb{U} \times \mathbb{R}$. In the light of the preceeding remarks, since we know that $\delta_n^{(u)} \in \langle \mathcal{C}\, u_n \rangle^\perp$, a natural initialization for such a gradient descent scheme would be

$$\bar{\omega}^0 := \mathcal{C}\, u_n - P_{\langle \mathcal{C}\, u_n \rangle} u_n \in \langle \mathcal{C}\, u_n \rangle^\perp, \quad \bar{\nu}^0 = 0.$$

Then $D(\bar{w}^0, \bar{\nu}^0)$ simplifies somewhat to

$$\begin{aligned} D(\bar{w}^0, \bar{\nu}^0) := -\nabla P(\bar{w}, \bar{\nu}) &= \begin{pmatrix} -\mathcal{M}_{\lambda_n}^* (\mathcal{M}_{\lambda_n} \bar{w}^0 - \bar{\nu} f_n + g_n) - s_n f_n \\ \langle f_n, \mathcal{M}_{\lambda_n} \bar{w} - \bar{\nu} f_n + g_n \rangle \end{pmatrix} \\ &= \begin{pmatrix} D_1(\bar{w}^0, \bar{\nu}^0) \\ D_2(\bar{w}^0, \bar{\nu}^0) \end{pmatrix}. \end{aligned} \tag{4.31}$$

The first component does perhaps not belong to $\langle \mathcal{C}\, u_n \rangle^\perp$ but is close to. This suggests taking

$$\begin{pmatrix} \bar{w}^1 \\ \bar{\nu}^1 \end{pmatrix} = \begin{pmatrix} \bar{w}^0 \\ \bar{\nu}^0 \end{pmatrix} + \xi \begin{pmatrix} P_{\langle \mathcal{C}\, u_n \rangle^\perp} D_1(\bar{w}^0, \bar{\nu}^0) \\ D_2(\bar{w}^0, \bar{\nu}^0) \end{pmatrix},$$

for a suitable step-size $\xi > 0$, and repeat the step based on (4.31). One then obtains analogous update formulas for $\begin{pmatrix} \bar{w}^k \\ \bar{\nu}^k \end{pmatrix}$. In fact, one could even determine (at the expense of further approximate applications of $\mathcal{C}, \mathcal{C}^*$) an optimal stepsize but we skip corresponding details.

Executing such a descent strategy requires only the scheme $[\mathcal{C}, f; \eta]$ (as well as a similar variant $[\mathcal{C}^*, f; \eta]$). Without going into further details (deferred to forthcoming numerical work) the update tolerances in those applications need to be fixed fractions of the target accuracy η_n in (4.18) whith ζ depending on which convergence strategy is being pursued, see (4.19) or (4.10).

Finally, whether the target accuracy has been met for a given iterate $\begin{pmatrix} \bar{w}^k \\ \bar{\nu}^k \end{pmatrix}$ can be checked by evaluating the residual $Q(\bar{w}^k, \bar{\nu}^k)$ from (4.29). Therefore, this strategy allows one, in principle, to realize the Newton update (4.1) in an accuracy controlled fashion. Due to the controlled condition of $\mathcal{M}_{\lambda_n}$ on $\langle \mathcal{C}\, u_n \rangle^\perp$ an error reduction by a fixed factor, required by (4.19), will be achieved after a uniformly bounded number of descent steps while strategy (4.10) would require the order of $|\log \bar{e}_n|$ steps.

As announced above, returning to (4.26), we briefly sketch now an alternate approach towards an error controlled approximation of $\mathcal{M}_{\bar{\lambda}}^{-1} \mathcal{C}\, \bar{u}$ when Δ is known and $(\bar{\lambda}, \bar{u})$ is already an accurate approximation to the principal eigenpair (λ°, u°). To that end, it will be convenient to rewrite

$$\overline{\Delta} = \mu^\circ |\nu - \mu^\circ|, \quad \mu^\circ = (\lambda^\circ)^{-1}, \nu = \arg\max\{|\mu| : \mu \in \sigma(\mathcal{C}) \setminus = \{\mu^\circ\}\}. \tag{4.32}$$

Specifically, assume that (e.g. by the above technique or with the aid of the scheme discussed in the next section) $\bar{\mu} = \bar{\lambda}^{-1}$ already satisfies $|\bar{\mu} - \mu^\circ| \leq \lambda^\circ \overline{\Delta}/8$, say.

Consider a disc $\Omega \subset \mathbf{C}$ with boundary Γ, center $\bar{\mu}$ and radius $\lambda^\circ\overline{\Delta}/2$. Thus Γ keeps a distance at least $3\lambda^\circ\overline{\Delta}/8$ from μ° and ν hence from $\sigma(\mathcal{C})$, while $|\Gamma| = \pi\lambda^\circ\overline{\Delta}$. Writing

$$\mathcal{M}_{\bar{\lambda}}^{-1}\,\mathcal{C} = \bar{\mu}(\bar{\mu}\,\mathcal{I} - \mathcal{C})^{-1}\,\mathcal{C},$$

classical functional calculus says that

$$\mathcal{M}_{\bar{\lambda}}^{-1}\,\mathcal{C}\,\bar{u} = \frac{1}{2\pi i}\int_\Gamma \frac{\bar{\mu}\zeta}{\bar{\mu}-\zeta}(\mathcal{I}\,\zeta - \mathcal{C})^{-1}\bar{u}d\zeta, \tag{4.33}$$

(see Sect. 3.2). By construction, the resolvent $(\zeta\,\mathcal{I} - \mathcal{C})^{-1}$ possesses a holomorphic extension to a symmetric strip A of width $3\lambda^\circ\Delta/8$ around Γ. It is well known (see e.g. [47]) that the trapezoidal rule provides then an approximation to the above contour integral with exponential accuracy. Specifically, consider equidistant quadrature points $\zeta_j \in \Gamma$, $j = 1, \ldots, N_\varepsilon$, and

$$I_h(\bar{u}) := \frac{h}{2\pi i}\sum_{j=1}^{\pi\lambda^\circ\Delta/h}\frac{\bar{\mu}\zeta_j}{\bar{\mu}-\zeta_j}(\mathcal{I}\,\zeta_j - \mathcal{C})^{-1}\bar{u}. \tag{4.34}$$

Then, employing $N = \pi\lambda^\circ\overline{\Delta}/h$ quadrature points (see e.g. [18, 33, 34])

$$\left\|I_h(\bar{u}) - \mathcal{M}_{\bar{\lambda}}^{-1}\,\mathcal{C}\,\bar{u}\right\|_{\mathbb{U}} \le Ce^{-\pi 3\lambda^\circ\overline{\Delta}/h8} = Ce^{-3N/8}.$$

Thus, given the exact states $(\mathcal{I}\,\zeta_j - \mathcal{C})^{-1}\bar{u}$, a constant multiple of $N_\varepsilon \eqsim |\log\varepsilon|$ quadrature points suffice to realize accuracy ε. Hence, if in addition one is able to generate approximations $r_j(\bar{u}) \approx (\mathcal{I}\,\zeta_j - \mathcal{C})^{-1}\bar{u}$ of order ε, i.e.,

$$\left\|I_{h_\varepsilon}(\bar{u}) - \frac{h_\varepsilon}{2\pi i}\sum_{j=1}^{N_\varepsilon}\frac{\bar{\mu}\zeta_j}{\bar{\mu}-\zeta_j}r_j(\bar{u})\right\|_{\mathbb{U}} \lesssim \varepsilon,$$

one generates approximations to $\mathcal{M}_{\bar{\lambda}}^{-1}\,\mathcal{C}\,\bar{u}$ of order ε. This leaves the task of approximately solving the $|\log\varepsilon|$ operator equations

$$(\zeta_j\,\mathcal{I} - \mathcal{C})w_j = \bar{u}, \quad j = 1, \ldots, N_\varepsilon. \tag{4.35}$$

The principal gain is that in these equations the ζ_j remain uniformly away from $\sigma(\mathcal{C})$ (with a distance of order $\mu^\circ\,\Delta$) and remain in this sense well conditioned. Moreover, (4.35) is equivalent to solving

$$(\mathcal{B} - \zeta_j^{-1}\,\mathcal{F})w_j = \zeta_j^{-1}\bar{u},$$

which has a similar structure as the original source problem, now with a modified global part, hence should be amenable to an adapted version of $[\mathcal{B}^{-1}, f; \eta]$. Details are left to forthcoming work.

5 Initialization Strategies and Power Iteration

A fully certified numerical realization of (CC1)–(CC3), based on Newton's method as idealized iteration, requires an initial guess in the admissible neighborhood specified in Theorem 4.1. First, this requires knowledge of the quantities $\|\mathcal{C}\|, \theta, \lambda^\circ, u^\circ$, which are generally not known. We recall that θ is small when the spectral gap Δ is small, see (3.26), (3.24). Therefore, the overarching objective in this section is to approximate these quantities, in principle, as accurately as one wishes, without the need of a sufficiently close initial guess. This calls for a globally convergent idealized iteration in (CC1).

A natural idea is then to employ a power iteration, as an idealized iteration in (CC1). To be specific, let $a_0^\circ \in \mathbb{U}$ with $\|a_0^\circ\| = 1, \alpha := \langle a_0^\circ, u_1\rangle \neq 0$ be an initial guess. Then, define

$$a_{n+1} := \mathcal{C}\, a_n^\circ, \quad a_{n+1}^\circ := \frac{a_{n+1}}{\|a_{n+1}\|}, \quad n = 0, 1, 2, \ldots. \tag{5.1}$$

The associated Rayleigh quotients are

$$\rho_{n+1} := \langle \mathcal{C}\, a_n^\circ, a_n^\circ\rangle, \quad n = 0, 1, 2, \ldots. \tag{5.2}$$

Again, in the spirit of previous discussions, we proceed in two stages: the first goal is to understand quantitative convergence characteristics of this idealized power method, formulated in this section, we focus on this task. With quantitative convergence results on these iterations at hand, one would contrive in a second step approximate numerical approximations of the idealized iteration obeying suitable accurracy tolerances that ensure convergence to the correct principal eigenpair. Since for the power-method this second step relies in essence on the ability to have an accuracy controlled application of $\mathcal{C}$ we dispense with corresponding details that in spirit follow the same lines as in previous sections but rather focus on the convergence of (5.1) and (5.2) in the infinite-dimensional setting.

The convergence of the power method for a general (non-normal) compact operator is already of interest in its own right since there does not seem to be much known, if anything at all. One reason is perhaps that the classical convergence proof for finite dimensional matrices is intrinsically finite dimensional and does not carry over to the infinite-dimensional case. For once, the concept of diagonalizability is far too restrictive. Moreover, for a fixed finite dimension an eigenbasis is automatically stable but its condition may depend on the dimension unless one is dealing with a normal matrix. For the present type of operator in infinite dimension, neither the notion of eigenbasis is appropriate nor are we lacking the necessary stability of an expansion system. The resulting main result reads as follows.

Theorem 5.1 *For $\mathcal{C}$, defined in (2.30) the power iteration (5.1), (5.2) converges linearly to the principal eigenpair $(u_1, \mu_1) \in \tilde{\mathbb{U}}$, i.e., for any $1 > \bar{\delta} > 1 - \Delta$ there exists a constant $C < \infty$ such the quantities a_n°, ρ_n, defined in (5.1) and (5.2), satisfy*

$$\|u_1 - a_n^\circ\| \leq C\bar{\delta}^n, \quad |\mu_1 - \rho_n| \leq C\bar{\delta}^n, \quad n \in \mathbb{N}. \tag{5.3}$$

The constant C depends on $\sigma(\mathcal{C})$ and may be arbitrarily large. If in addition one assumes $\mathcal{C}$ belongs to the Schatten class S_p of compact operators on $\mathbb{U}$ for some $p < \infty$, the constant C in (5.3) can be quantified in terms of a_0°, $\|\mathcal{C}\|_p$, and Δ.

For the convenience of the reader we recall below the precise definition of the Schatten class S_p and a few known relevant facts. Membership to a Schatten class could perhaps be viewed as replacing the knowledge of the condition of an eigenbasis in the matrix case.

As a substitute for eigen-expansions in the matrix case we resort to classical functional calculus and Riesz projections, see Sect. 3.2.

Exploiting the bounds (5.3) for step (CC2), i.e., for an accuracy controlled approximate execution of (5.1) and (5.2) not only requires an estimate of the constant C in (5.3) but also a quantitative (hopefully sharp) lower bound of the spectral gap Δ. The importance of knowing such an estimate has already transpired in previous sections.

In summary, it seems that intrinsic difficulties of numerically solving spectral problems with certified accuracy for the current scope of unsymmetric compact operators can be reduced to some extent to assessing Δ (and this fact is actually confirmed by works such as [5]). Therefore, we highlight in Sect. 5.2 the difficulties in estimating Δ and the role of heuristics in gaining computational information on Δ.

The statements in Theorem 5.1 reflect that we are not able to quantify convergence under the mere assumption of compactness when $\mathcal{C}$ is not normal. As we will see later below the difficulty is to quantify resolvent bounds of the type $\max_{\zeta\in\Gamma}\|\mathcal{R}_{\mathcal{C}}(\zeta)\|$ in terms of the distance of the contour Γ from the spectrum. In fact, in the most favorable situation that $\mathcal{C}$ is a normal operator, it is known that

$$\|\mathcal{R}_{\mathcal{C}_\omega}(\zeta)\| \le \big(\inf_{\zeta'\in\Gamma_\omega}|\zeta-\zeta'|\big)^{-1} := d(\zeta,\Gamma_\omega)^{-1}, \tag{5.4}$$

while it is known that (5.4) is generally not true when $\mathcal{C}$ is not normal. To see this, we recall the following version of Schur's Lemma that ensures for every compact operator $\mathcal{A}$ the existence of a decomposition $\mathcal{A} = D + N$ ([7, Theorem 3.2]) where D, N are compact, D is normal $\sigma(\mathcal{A}) = \sigma(D)$, and N is quasi-nilpotent, i.e., $\sigma(N) = \{0\}$. Then powers of $\mathcal{A}$ are of the form

$$\mathcal{A}^k = D^k + B_k,$$

where B_k is a polynomial in D and N of degree k. Although, by compactness, properly rescaled powers of N will (eventually) tend to zero it is not so clear what the right scaling for the mixed terms in B_k should be in order to produce a quantitative damping of B_k as k tends to infinity. In fact, the deviation of normality of $\mathcal{A}$ reflected by N will play later a role in the proof of Theorem 5.1.

The deviation from normality can indeed be quantified when $\mathcal{C}$ belongs to some Schatten class. We briefly recall this concept for a general compact operator $\mathcal{A}$ on a Hilbert space H. Let S_∞ denote the set of all compact operators on H. Then the p-th Schatten class is defined as (see e.g. [7, 22])

$$S_p := \Big\{ \mathcal{A} \in S_\infty : \| \mathcal{A} \|_p := \Big(\sum_{j=1}^{\infty} s_j(\mathcal{A})^p \Big)^{1/p} < \infty \Big\},$$

where the $s_j(\mathcal{A})$ are the singular values of $\mathcal{A}$. Obviously, every finite rank operator belongs to every S_p, $p \in (0, \infty]$. In general, the larger p the weaker the condition.

For later use in Sect. 5.1 we invoke now some results from [7] which we briefly formulate again in general terms for a general compact operator $\mathcal{A}$ on a Hilbert space H. Recall the Schur decomposition $\mathcal{A} = D+N$ where D is normal with $\sigma(\mathcal{A}) = \sigma(D)$ and N quasi-nilpotent. Since $(z\mathcal{I} - \mathcal{A})^{-1} = (\mathcal{I} - (z\mathcal{I} - D)^{-1}N)^{-1}(z\mathcal{I} - D)^{-1}$ one gets

$$\begin{aligned} \|(z\mathcal{I} - \mathcal{A})^{-1}\| &\le \|(\mathcal{I} - (z\mathcal{I} - D)^{-1}N)^{-1}\| \|(z\mathcal{I} - D)^{-1}\| \\ &\le \|(\mathcal{I} - (z\mathcal{I} - D)^{-1}N)^{-1}\| d(z, \sigma(D))^{-1}, \end{aligned}$$

ending up with the task of bounding $\|(\mathcal{I} - (zI - D)^{-1}N)^{-1}\|$. This is done in [7, Theorem 4.1] providing for $\mathcal{A} \in S_p$, $p < \infty$,

$$\|(z\mathcal{I} - \mathcal{A})^{-1}\| \le d(z, \sigma(\mathcal{A}))^{-1} \exp\Big\{ \frac{a_p 2^{1+(p-1)_+} \| \mathcal{A} \|_p^p}{d(z, \sigma(\mathcal{A}))^p} + b_p \Big\}, \tag{5.5}$$

a_p, b_p constants depending only on p.

Remark 5.2 (see [7, Remark 2.2]) For certain values of p the constants a_p, b_p are known:

1. $p = 1$: $a_1 = 1$, $b_1 = 0$;
2. $p = 2$: $a_2 = \frac{1}{2} = b_2$.
3. For $0 < p \le 1$ a possible choice is

$$a_p = \sup_{z \in \mathbf{C}} |z|^{-p} \log |(1+z)|, \quad b_p = 0, \tag{5.6}$$

while for $p > 1$ one can take for a_p any real number larger than

$$\underline{p} := \sup_{z \in \mathbf{C}} \log \Big| (1+z) \exp\Big\{ \sum_{j=1}^{\lceil p \rceil - 1} \frac{(-z)^j}{j} \Big\} \Big|, \tag{5.7}$$

with $b_p \neq 0$ depending on a_p.

5.1 Proof of Theorem 5.1

We take up again the notation from Sect. 3.2 for the spectrum

$$\sigma(\mathcal{C}) = \{\mu_j : j = 1, \dots, \infty\},$$

where $|\mu_j|$ decreases with increasing $j \in \mathbb{N}$ (recall the correspondence $\mu_1 = \mu^\circ = (\lambda^\circ)^{-1}$, $u^\circ = u_1$). Note that the spectral gap (3.24) can equivalently be written as

$$\overline{\Delta} = \left|1 - \frac{\mu_2}{\mu_1}\right| \geq 1 - \frac{|\mu_2|}{\mu_1} = \Delta.$$

Finally, it will be convenient to assume in what follows the normalization $\|u_1\| = 1$ for the principal eigenstate u_1.

The proof of Theorem 5.1 is based on analyzing the action of $\mathcal{C}$ on invariant subspaces of the form $V_\omega := \mathcal{E}_\mathcal{C}(\omega)\,\mathbb{U}$ for subsets $\omega \subset \sigma(\mathcal{C})$. The following facts are probably standard but we include them for completeness.

Lemma 5.3 *(a) For $\omega \subset \sigma(\mathcal{C})$ let $\mathcal{C}_\omega := \mathcal{C}\,|_{V_\omega}$ denote the restriction of $\mathcal{C}$ to V_ω (recall that $\mathcal{C}\,V_\omega = V_\omega$). We have*

$$\mathcal{R}_\mathcal{C}(\zeta)\,|_{V_\omega} = \mathcal{R}_{\mathcal{C}_\omega}(\zeta), \tag{5.8}$$

and the representation

$$\mathcal{C}_\omega^\ell = \frac{1}{2\pi i}\int_{\Gamma_\omega} \zeta^\ell\, \mathcal{R}_{\mathcal{C}_\omega}(\zeta)\, d\zeta, \tag{5.9}$$

holds for every $\ell \in \mathbb{N}$ where, as before, $\Gamma_\omega = \partial\Omega$ for some domain $\Omega \subset \mathbf{C}$ such that $\omega \subset \Omega$, $\Omega \cap (\sigma(\mathcal{C})\backslash\omega) = \emptyset$. (b) Moreover, we have

$$\|\,\mathcal{C}^\ell\, u\| \leq \frac{|\Gamma_\omega|}{2\pi}\max_{\zeta\in\Gamma_\omega} |\zeta|^\ell \|\,\mathcal{R}_{\mathcal{C}_\omega}(\zeta)u\|, \quad \forall\, u \in V_\omega, \tag{5.10}$$

where $|\Gamma_\omega|$ denotes the length of Γ_ω.

Proof By the projection property of $\mathcal{E}_\mathcal{C}$ we have for every $u \in V_\omega$ that

$$u = \mathcal{E}_\mathcal{C}(\omega)u = \frac{1}{2\pi i}\int_{\Gamma_\omega} \mathcal{R}_\mathcal{C}(\zeta)u\, d\zeta.$$

When $|\zeta| > \|\,\mathcal{C}\,\| \geq \|\,\mathcal{C}_\omega\,\|$ we have for any $u \in V_\omega$

$$\begin{aligned}(\zeta\,\mathcal{I} - \mathcal{C}_\omega)^{-1}u &= \zeta^{-1}(\mathcal{I} - \zeta^{-1}\,\mathcal{C}_\omega)^{-1}u = \zeta^{-1}\sum_{j=0}^{\infty}\frac{\zeta^{-j}\,\mathcal{C}_\omega^j\,\mathcal{E}_\mathcal{C}(\omega)^j u}{j!}\\ &= \zeta^{-1}\sum_{j=0}^{\infty}\frac{\zeta^{-j}\,\mathcal{C}^j\,u}{j!} = (\zeta\,\mathcal{I} - \mathcal{C})^{-1}u.\end{aligned} \tag{5.11}$$

Since the resolvent is holomorphic on $\rho(\mathcal{C})$ the above agreement must hold for all $\zeta \in \rho(\mathcal{C})$, confirming (5.8).

Clearly, $\sigma(\mathcal{C}_\omega) = \omega$. Let Ω'_ω with boundary Γ' be any domain in $\mathbf{C}$ containing Γ_ω while one still has $\Omega'_\omega \cap (\sigma(\mathcal{C}) \setminus \omega) = \emptyset$. Then, we have (by standard holomorphic

functional calculus) $\mathcal{C}_\omega^\ell = \frac{1}{2\pi i}\int_{\Gamma'} \xi^\ell\, \mathcal{R}_{\mathcal{C}_\omega}(\xi)\, d\xi$. Next recall Hilbert's formula (1st resolvent formula)

$$\mathcal{R}_{\mathcal{C}}(\zeta)\mathcal{R}_{\mathcal{C}}(\xi) = (\xi-\zeta)^{-1}\big(\mathcal{R}_{\mathcal{C}}(\zeta) - \mathcal{R}_{\mathcal{C}}(\xi)\big). \tag{5.12}$$

Since Ω'_ω strictly contains Γ_ω we obtain for $u \in V_\omega$ (so that $u = \mathcal{E}_{\mathcal{C}}(\omega)u$)

$$\begin{aligned}\mathcal{C}_\omega^\ell\, u &= \Big(\frac{1}{2\pi i}\Big)^2 \int_{\Gamma'_\omega}\int_{\Gamma_\omega} \xi^\ell\, \mathcal{R}_{\mathcal{C}}(\xi)\, \mathcal{R}_{\mathcal{C}}(\zeta) u\, d\xi\, d\zeta \\ &= \Big(\frac{1}{2\pi i}\Big)^2 \int_{\Gamma'_\omega}\int_{\Gamma_\omega} \frac{\xi^\ell}{\zeta-\xi}\big(\mathcal{R}_{\mathcal{C}}(\xi) - \mathcal{R}_{\mathcal{C}}(\zeta)\big) u\, d\xi\, d\zeta.\end{aligned}$$

Since residue calculus yields

$$\frac{1}{2\pi i}\int_{\Gamma'_\omega} \frac{\xi^\ell}{\zeta-\xi}\, \mathcal{R}_{\mathcal{C}}(\zeta) u\, d\xi = -\zeta^\ell\, \mathcal{R}_{\mathcal{C}}(\zeta) u, \quad \frac{1}{2\pi i}\int_{\Gamma_\omega} \frac{\xi^\ell}{\zeta-\xi}\, \mathcal{R}_{\mathcal{C}}(\xi) u\, d\zeta = 0,$$

(the second relation, because ξ is outside Ω_ω) relation (5.9) follows. The rest of the assertion (b) follows from (5.9). □

To see whether $(\mu_1^{-1}\,\mathcal{C})^\ell$ contracts on subspaces V_ω when $\mu_1 \notin \omega$, we specialize Lemma 5.3 to a domain $\Omega = \Omega_\varepsilon$ with boundary Γ_ε such that

$$\mu_1 \notin \Omega_\varepsilon,\ \ \sigma_{>1} := \sigma(\mathcal{C}) \setminus \{\mu_1\} \subset \Omega_\varepsilon,\ \ \operatorname{dist}(\Gamma_\varepsilon, \sigma_{>1}) \geq \operatorname{dist}(\Gamma_\varepsilon, \mu_2) = \varepsilon. \tag{5.13}$$

Here ε is a fixed constant whose value will be stipulated later. The goal is to show that the scaled powers $(\mu_1^{-1}\,\mathcal{C})^\ell$ (eventually) contract on the invariant subspaces $V_{\sigma_{>1}}$.

Lemma 5.4 *Adhering to the above notation, let* $1 - \Delta = \mathfrak{q} < \overline{\delta} < 1$. *Then, there exists an* $\ell_0 \in \mathbb{N}$ *such that*

$$\|(\mu_1^{-1}\,\mathcal{C})^\ell \mathcal{E}_{\mathcal{C}}(\sigma_{>1})u\| \leq \overline{\delta}^{\ell-\ell_0}\|\mathcal{E}_{\mathcal{C}}(\sigma_{>1})u\|, \quad \ell \in \mathbb{N},\ u \in \mathbb{U}. \tag{5.14}$$

If one assumes in addition that $\mathcal{C} \in \mathcal{S}_p$ *for some* $1 \leq p < \infty$, *then* $\ell_0 \in \mathbb{N}$ *can be bounded in terms of* $\overline{\delta}$, $\|\,\mathcal{C}\,\|_p$, Ω_ε.

Proof It is well-known that $\mathcal{R}_{\mathcal{C}}(\zeta)$ is bounded for each $\zeta \notin \sigma(\mathcal{C})$. Hence $M_\varepsilon := \max_{\zeta\in\Gamma_\varepsilon} \|\mathcal{R}_{\mathcal{C}}(\zeta)\| < \infty$. Thus, (5.10) yields for any $u \in V_{\sigma_{>1}}$

$$\|(\mu_1^{-1}\,\mathcal{C})^\ell u\| \leq \frac{|\Gamma_\varepsilon|}{2\pi}\Big(\frac{\max_{\zeta\in\Gamma_\varepsilon}|\zeta|}{\mu_1}\Big)^\ell M_\varepsilon\|u\| \leq \frac{|\Gamma_\varepsilon|}{2\pi}\Big(\frac{|\mu_2|+\varepsilon}{\mu_1}\Big)^\ell M_\varepsilon\|u\|, \quad u \in V_{\sigma_{>1}}. \tag{5.15}$$

Now recall that $\Delta = 1 - \frac{|\mu_2|}{\mu_1}$ and fix some $\beta \in (0,1)$ to define

$$\varepsilon := \beta(\mu_1 - |\mu_2|) = \beta\mu_1\Delta, \quad \overline{\delta} := \frac{|\mu_2|+\varepsilon}{\mu_1} = \Delta + \beta(1-\Delta) = \beta + (1-\beta)\Delta. \tag{5.16}$$

Thus, $\overline{\delta}$ is strictly less than one for $\beta < 1$ and tends to Δ when β tends to zero. Hence, on account of (5.15), (5.14) follows with

$$\ell_0 := \min\left\{\ell \in \mathbb{N} : \frac{\overline{\delta}^\ell |\Gamma_\varepsilon| M_\varepsilon}{2\pi} \le 1\right\}. \tag{5.17}$$

This confirms the first part of the assertion. □

Remark 5.5 For a general compact non-normal operator $\mathcal{C}$ on $\mathbb{U}$ we cannot further quantify the constant $M_\varepsilon = \max_{\zeta\in\Gamma_\varepsilon} \|\mathcal{R}_\mathcal{C}(\zeta)\|$ which could be arbitrarily large. Hence, ℓ_0 in (5.17) could be arbitrarily large which prevents us from concluding any meaningful bound on the constant in (5.3).

Now assume that in addition $\mathcal{C} \in S_p$ for some $1 \le p < \infty$. Then (5.10) combined with (5.5) yields

$$\begin{aligned}\|\mathcal{C}^\ell u\| &= \|\mathcal{C}^\ell_{\sigma_{>1}} u\| \\ &\le \frac{|\Gamma_\varepsilon|}{2\pi}(|\mu_2| + \varepsilon)^\ell \varepsilon^{-1} \exp\left\{\frac{2^{1+(p-1)_+} a_p \|\mathcal{C}_{\sigma_{>1}}\|_p^p}{\varepsilon^p} + b_p\right\}\|u\|, \quad u \in V_{\sigma_{>1}}.\end{aligned} \tag{5.18}$$

Next observe that $\mathcal{C}_{\sigma_{>1}} = \mathcal{C}\, E_\mathcal{C}(\sigma_{>1})$ belongs to S_p when $\mathcal{C}$ does. In fact,

$$\|\mathcal{C}_{\sigma_{>1}}\|_p \le 2\|\mathcal{C}\|_p, \tag{5.19}$$

To see this, recall that $\mathcal{C} = \mathcal{C}\, E_\mathcal{C}(\mu_1) + \mathcal{C}\, E_\mathcal{C}(\sigma_{>1})$ so that $\|\mathcal{C}\, E_\mathcal{C}(\sigma_{>1})\|_p \le \|\mathcal{C}\|_p + \|\mathcal{C}\, E_\mathcal{C}(\mu_1)\|_p$. Since $\mathcal{C}\, E_\mathcal{C}(\mu_1)$ has rank one its single non-zero singular value is given by the square root of

$$\max_{u\in\mathbb{U}} \frac{\langle E_\mathcal{C}(\mu_1)u, \mathcal{C}\, E_\mathcal{C}(\mu_1)(u)\rangle}{\|u_1\|^2} = \frac{\langle E_\mathcal{C}(\mu_1)u_1, \mathcal{C}\, E_\mathcal{C}(\mu_1)u_1\rangle}{\|u_1\|^2} = \mu_1^2.$$

Since by Weyl's inequality eigenvalues are dominated by singular values (5.19) follows.

Hence, we conclude from (5.18) that

$$\|\mu_1^{-\ell}\mathcal{C}^\ell_{\sigma_{>1}} u\| \le \left(\frac{|\mu_k| + \varepsilon}{\mu_1}\right)^\ell M(\beta, \Delta, \mathcal{C}, p)\|u\|, \quad u \in V_{\sigma_{>1}}, \tag{5.20}$$

where

$$M(\beta, \Delta, \mathcal{C}, p) := \frac{|\Gamma_\varepsilon|}{2\pi\varepsilon} \exp\left\{\frac{2^{p+1+(p-1)_+} a_p \|\mathcal{C}\|_p^p}{(\beta(\mu_1 - |\mu_2|))^p} + b_p\right\}. \tag{5.21}$$

Taking in analogy to (5.17)

$$\ell_0 = \ell_0(\Delta, \beta, \mathcal{C}, p) := \operatorname{argmin}\big\{\ell \in \mathbb{N} : \overline{\delta}^{\ell} M(\Delta, \beta, \mathcal{C}, p) \le 1\big\}, \tag{5.22}$$

finishes the proof.

To proceed we define a new (equivalent) norm on $\mathbb{U}$ by

$$|\!|\!| u |\!|\!| := \|E_{\mathcal{C}}(\mu_1)u\| + \|E_{\mathcal{C}}(\sigma_{>1})u\|, \tag{5.23}$$

i.e., there exist constants c_1, C_1, depending on $\mathcal{C}$ such that

$$c_1 |\!|\!| \cdot |\!|\!| \le \|\cdot\| \le C_1 |\!|\!| \cdot |\!|\!|. \tag{5.24}$$

Recalling the shorthand notation $\langle u\rangle := \operatorname{span}\{u\}$, it will be convenient to introduce the "distance"

$$\operatorname{dist}(\langle u\rangle, \langle u_1\rangle) := \min_{c\in\mathbb{R}} \|cu - u_1\|. \tag{5.25}$$

Lemma 5.6 *Assume that $a_0^\circ \in \mathbb{U}$ satisfies $\|a_0^\circ\| = 1$ and $a_0^\circ \ge 0$ so that*

$$E_{\mathcal{C}}(\mu_1)a_0^\circ = \alpha u_1 \quad \textit{for some } \alpha > 0.$$

Let a_n°, ρ_n, ℓ_0 be defined by (5.1), (5.2), and (5.22). Then we have

$$\alpha \operatorname{dist}(\langle a_n^\circ\rangle, \langle u_1\rangle) \le \|(\mu_1^{-1}\mathcal{C})^n a_0^\circ - \alpha u_1\| \le \frac{C_1}{c_1}\overline{\delta}^{n-\ell_0}, \quad n \in \mathbb{N}. \tag{5.26}$$

Furthermore,

$$\|a_n^\circ - u_1\| \le \sqrt{2}\operatorname{dist}(\langle a_n^\circ\rangle, \langle u_1\rangle) \le \frac{2C_1}{\alpha c_1}\overline{\delta}^{n-\ell_0}, \quad n \in \mathbb{N}, \tag{5.27}$$

and

$$|\rho_{n+1} - \mu_1| \le C\overline{\delta}^{n-\ell_0}, \quad n \in \mathbb{N}, \tag{5.28}$$

where $C = \alpha^{-1}(\mu_1 + \|\mathcal{C}\|)\frac{C_1}{c_1}$.

Proof Since $u_1 = (\mu_1^{-1}\,\mathcal{C})^n u_1$ and $E_{\mathcal{C}}(\sigma_{>1})(\mu_1^{-1}\,\mathcal{C})^n u_1 = 0$ for all $n \in \mathbb{N}$, we have

$$\begin{aligned}
C_1^{-1}\|(\mu_1^{-1}\,\mathcal{C})^n a_0^\circ - \alpha u_1\| &\le |\!|\!|(\mu_1^{-1}\,\mathcal{C})^n a_0^\circ - \alpha u_1|\!|\!| \\
&= \|E_{\mathcal{C}}(\mu_1)((\mu_1^{-1}\,\mathcal{C})^n a_0^\circ - \alpha u_1)\| + \|E_{\mathcal{C}}(\sigma_{>1})(\mu_1^{-1}\,\mathcal{C})^n a_0^\circ\| \\
&= \|((\mu_1^{-1}\,\mathcal{C})^n (E_{\mathcal{C}}(\mu_1)a_0^\circ - \alpha u_1)\| + \|E_{\mathcal{C}}(\sigma_{>1})(\mu_1^{-1}\,\mathcal{C})^n a_0^\circ\| \\
&= \|(\mu_1^{-1}\,\mathcal{C})^n (E_{\mathcal{C}}(\sigma_{>1})a_0^\circ)\| \\
&\le \overline{\delta}^{n-\ell_0}\|E_{\mathcal{C}}(\sigma_{>1})a_0^\circ\| \\
&\le \overline{\delta}^{n-\hat{\ell}}|\!|\!|a_0^\circ|\!|\!| \le c_1^{-1}\overline{\delta}^{n-\hat{\ell}}\|a_0^\circ\| = c_1^{-1}\overline{\delta}^{n-\hat{\ell}},
\end{aligned}$$

where we have use Lemma 5.4 in the second but last line. This confirms the second inequality in (5.26). Since $\langle a_n^\circ\rangle = \langle \mathcal{C}^n\, a_0^\circ\rangle = \langle c\,\mathcal{C}^n\, a_0^\circ\rangle$ for all $c \neq 0$ and $n \in \mathbb{N}$, the first inequality in (5.26) follows as well.

To proceed let $s_n := \operatorname{argmin}_{s\in\mathbb{R}}\|s a_n^\circ - u_1\| = \langle a_n^\circ, u_1\rangle$. Since $a_0^\circ \ge 0$ and $\mathcal{C}$ is a positive operator we have $s_n \ge 0$ for all $n \in \mathbb{N}$. Observing by direct calculation that $\|a_n^\circ - u_1\|^2 = 2(1 - s_n)$ and $\|s_n a_n^\circ - u_1\|^2 = 1 - s_n^2$, we obtain

$$\operatorname{dist}\big(\langle a_n^\circ\rangle, \langle u_1\rangle\big)^2 = \|s_n a_n^\circ - u_1\|^2 = (1 + s_n)(1 - s_n) = \frac{1 + s_n}{2}\|a_n^\circ - u_1\|^2,$$

which, upon using (5.26), proves (5.27).

Concerning (5.28), one has

$$\begin{aligned}
|\rho_{n+1} - \mu_1| = |\langle \mathcal{C}\, a_n^\circ, a_n^\circ\rangle - \langle \mathcal{C}\, u_1, u_1\rangle| &\le |\langle \mathcal{C}(a_n^\circ - u_1), a_n^\circ\rangle| + |\langle \mathcal{C}\, u_1, a_n^\circ - u_1\rangle| \\
&\le \|\,\mathcal{C}(a_n^\circ - u_1)\| + \mu_1\|a_n^\circ - u_1\| \le (\mu_1 + \|\,\mathcal{C}\,\|)\|a_n^\circ - u_1\|
\end{aligned}$$

which completes the proof. □

5.2 Practical Aspects

The preceding developments neither pretend nor intend to offer a directly applicable algorithm, but rather formulate concepts that, in contrast to the "classical" approach realize accuracy quantification without (unrealistic) regularity assumptions. Nevertheless, the various routines are implementable. For the work horse $[\mathcal{B}^{-1}, f; \eta]$ this has been demonstrated in [15]. A complete certifiable eigensolver based on the proposed paradigm, would of course require deviating significantly from existing software structures, hence requires in essence coding from scratch. More importantly, certifiability would require knowledge of (or at least good estimates for) the "problem parameters" $\|\,\mathcal{C}\,\|, \theta, \Delta$ which is generally not available. The question remains, are these quantities computationally accessible if the knowledge of specific optical parameters in the Boltzmann operator doesn't suffice. Regarding $\|\,\mathcal{C}\,\|$ an answer should be affirmative. In fact, the Rayleigh quotients for the self adjoint positive definite operator $\mathcal{C}^*\,\mathcal{C}$ would tend to $\|\,\mathcal{C}\,\|^2$ and the power method is in this case much easier to analyze and faster,

due to the underlying orthonormal eigensystem. This also yields an upper bound for μ°.

Assessing the quantities θ, Δ instead seems to be much harder. Although we do not know of a concise algorithmic strategy for approximating these quantities with quantifiable certainty, there are heuristics that we expect to provide sufficient information to underpin an implementation. This is beyond the scope (and also spirit) of the present paper so that we are content with some brief indications and leave a more careful elaboration to future work with a numerical focus. First, in view of the relations (3.26), a good (lower bound for Δ would also yield an indication of θ (up to a factor depending on the angle between eigenvector and singular vector in $\mathbb{U}_\circ^\perp$). Approximating θ directly, based on (3.20), could be attempted through a power method where one has to deal though with the restriction to $\mathbb{U}_\circ^\perp$ in each step. Although we don't know u° yet, in view of $\mathbb{U}_\circ^\perp = \mathcal{E}_{\mathcal{C}^*}(\sigma(\mathcal{C}^*) \setminus \{\mu^\circ\})\,\mathbb{U}$, this restriction can be realized through the Riesz projection $\mathcal{E}_{\mathcal{C}^*}(\sigma(\mathcal{C}^*) \setminus \{\mu^\circ\})$ which, in turn, can be approximated, in principle, by the quadrature approach, outlined at the end of Sect. 4.4. However, the accuracy of this step is uncertain, precisely because we don't know Δ. Nevertheless, a lower bound for μ° might lead to a reasonable contour that permits further computational exploration. Such a lower bound, in turn, could be obtained from the Rayleigh quotients of the power iteration discussed above. Lacking knowledge of Δ, it seems that all one can do is to apply $\mathcal{C}$ with very high accuracy and monitor the increase of the Rayleigh quotients, admittedly, a heuristic approach whose cost to success cannot be estimated.

These comments suffice perhaps to appraise remaining intrinsic obstructions reflecting the principal hardness of this type of spectral problems.

6 Concluding Remarks

Our overarching goal is to develop and analyze a framework that eventually leads to an accuracy controlled solution of the criticality problem for a model version of the neutronic equation which nevertheless is general enough to exhibit its intrinsic difficulties. On the one hand, this model is behind important applications related to future energy production. On the other hand, it is an example of an intrinsically difficult spectral problem and many of the proposed concepts have some bearing beyond the specific model setting. Rendering a rigorous accuracy control, independent of any unrealistic a priory regularity requirements, is achieved by a paradigm that deviates strongly from common practice. It requires properly intertwining idealized iterations in a model compliant function space with dynamically updated numerical approximations that are controlled by a posteriori quantities. At no stage will the numerical outcome result from a single a priori chosen discretization. In both regards, stable variational formulations of the underlying PDE model play a key role. This paradigm, albeit in different formal disguises, has led in the past to a first complete complexity and convergence analysis of wavelet methods for PDEs [13], as well as of finite element methods [9], and of low-rank and tensor methods for high-dimensional PDEs and Uncertainty Quantification [8]. From an analytical perspective, the problem types in these latter scenarios are in many ways more benign. In the present setting it seems

that we cannot completely ensure convergence success, at least not in conjunction with quantitative complexity bounds, see the comments in Sect. 5.2. Just being able to computationally approximate the principal eigenpair within a desired accuracy is a challenge and the importance of the application may welcome even enormous computational effort if it leads to a certifiable improved outcome quality. In this regard, we have tried to complement the rigorous part of the analysis by possible computational techniques that better and better cope with the most essential knowledge gap, namely a sufficiently good estimate of the spectral gap.

Appendix A Proof of Lemma 2.4

We observe first that $\int_{\mathrm{D}\times\mathrm{V}} u(v\cdot\nabla u)\mathrm{d}x\mathrm{d}v \geq 0$ holds for every $u \in H_{0,-}(\mathrm{D}\times\mathrm{V})$ where $H_{0,-}(\mathrm{D}\times\mathrm{V})$ is defined in analogy to (2.20). In fact, integrating by parts gives

$$\int_{\mathrm{D}\times\mathrm{V}} u(v\cdot\nabla u)\mathrm{d}x\mathrm{d}v = -\int_{\mathrm{D}\times\mathrm{V}} u(v\cdot\nabla u)\mathrm{d}x\mathrm{d}v + \int_{\Gamma} u^2 v\cdot n.$$

Since $u \in H_{0,-}(\mathrm{D}\times\mathrm{V})$, this provides

$$\int_{\mathrm{D}\times\mathrm{V}} u(v\cdot\nabla u)\mathrm{d}x\mathrm{d}v = \frac{1}{2}\int_{\Gamma} u^2 v\cdot n = \frac{1}{2}\int_{\Gamma_+} u^2 v\cdot n \geq 0, \tag{A1}$$

by definition of Γ_+, see also [19, 28, p1105].

Next, given $g \in L_2(\mathrm{D}\times\mathrm{V})$, there exists a unique $u \in H_{0,-}(\mathrm{D}\times\mathrm{V})$ (see Theorem 2.1, (2.17)) such that $\mathcal{T}u = \mathcal{K}g$. Multiplying $\mathcal{T}u = \mathcal{K}g$ by u, integrating over $\mathrm{D}\times\mathrm{V}$, and using (A1) and $\kappa \geq 0$, we obtain

$$\begin{aligned}\int_{\mathrm{D}\times\mathrm{V}} \sigma u^2 &\leq \int_{\mathrm{D}\times\mathrm{V}\times\mathrm{V}} \kappa(x,v,v')u(x,v)g(x,v')\mathrm{d}x\mathrm{d}v\mathrm{d}v' \\ &\leq \int_{\mathrm{D}} \left(\int_{\mathrm{V}\times\mathrm{V}} \kappa(x,v,v')u(x,v)^2\mathrm{d}v\mathrm{d}v'\right)^{1/2} \\ &\quad \left(\int_{\mathrm{V}\times\mathrm{V}} \kappa(x,v,v')g(x,v')^2\mathrm{d}v\mathrm{d}v'\right)^{1/2}\mathrm{d}x,\end{aligned} \tag{A2}$$

where we have used Cauchy-Schwarz' inequality.

One deduces now from condition (H4) that $\int_{\mathrm{V}} \kappa(x,v,v')dv' \leq \rho\sigma(x,v)$ and $\int_{\mathrm{V}} \kappa(x,v,v')dv \leq \rho\sigma(x,v')$ for some constant $\rho < 1$. Using these bounds in the above inequality, yields

$$\int_{\mathrm{D}\times\mathrm{V}} \sigma|u|^2\mathrm{d}x\mathrm{d}v \leq \rho^2 \int_{\mathrm{D}\times\mathrm{V}} \sigma|g|^2\mathrm{d}x\mathrm{d}v$$

which means $\|u\|_{\mathbb{U}(\sigma)} \leq \rho\|g\|_{\mathbb{U}(\sigma)}$ and hence (2.25). □

Appendix B Proof of Lemma 2.7

One shows first that u° is strictly positive except on Γ_-. In fact, by the preceding observations,

$$\mu^\circ u^\circ = (\mathcal{I} - \mathcal{T}^{-1}\mathcal{K})^{-1}\mathcal{T}^{-1}\mathcal{F}u^\circ \geq \mathcal{T}^{-1}\mathcal{F}u^\circ.$$

Defining for $x, x' \in \mathrm{D}$ the optical path

$$p(x, x', v) := \int_0^{|x-x'|} \sigma\Big(x + s\frac{x'-x}{|x'-x|}, v\Big)ds,$$

and characteristic distance $d(x, v)$ from the boundary $\partial\,\mathrm{D}$, i.e., $x - d(x, v)v \in \partial\,\mathrm{D}$, the explicit representation of $\mathcal{T}^{-1}$ yields

$$\mu^\circ u^\circ(x, v) \geq \int_0^{d(x,v)} \exp\big(-p(x, x - sd(x, v), v)\big)(\mathcal{F}u^\circ)(x - sv, v)ds. \quad \text{(B1)}$$

One needs to show then that the right-hand side of (B1) is strictly positive except for $(x, v) \in \Gamma_-$. This is not quite obvious yet because a non-trivial u° could have the property that for x in a small neighborhood of some x_0, $u^\circ(x - sv, v') = 0$ holds for all s and $v' \in \mathrm{V}$. Therefore, $\mathcal{F}$ being linear and positive, an additional application of $\mathcal{F}$ to both sides of (B1) yields $\mu^\circ \mathcal{F}u^\circ \geq \mathcal{F}\mathcal{T}^{-1}\mathcal{F}u^\circ$. Thus, as soon as one shows $\mathcal{F}\mathcal{T}^{-1}\mathcal{F}u^\circ$ is strictly positive on $\overline{(\mathrm{D}\times\mathrm{V})\setminus\Gamma_-}$, corresponding strict positivity follows from (B1). To that end, elementary manipulations show that

$$(\mathcal{F}\mathcal{T}^{-1}\mathcal{F}u^\circ)(x, v) = \int_{\mathrm{V}} \varphi(x, v', v) \int_0^{d(x,v')} e^{-p(x,x-sv',v')} \int_{\mathrm{V}} \varphi(x - sv', v'', v')dv'u^\circ(x - sv', v'')\,dsdv'dv''.$$

In [19, Theorem 7, p. 1154] it is proved that the right-hand side expression is strictly positive in $\mathrm{D}\times\mathrm{V}$ except on Γ_-. We briefly explain the essence of the argument. By (H1) and (H2), D is convex, hence star-shaped for every $x \in \mathrm{D}$, and V contains a set that is homeomorphic to the sphere. Thus, the arguments $(x - sv', v'')$ of u° traverse all of $\mathrm{D}\times\mathrm{V}$. Moreover by (H5), $u^\circ(x - sv', v'')$ is multiplied by strictly positive quantities. Thus, for the right-hand side to vanish on a set of strictly positive measure in $(\mathrm{D}\times\mathrm{V})\setminus\Gamma_-$, u° would have to vanish everywhere in $\mathrm{D}\times\mathrm{V}$. This means, in particular, that $\mathcal{F}\mathcal{T}^{-1}\mathcal{F}$ can be represented as an integral operator with a strictly positive kernel. We refer to [19, Theorem 7, p. 1154] for detailed arguments to that end.

As a next step it is shown that u° is the only positive eigenstate. This is based on considering the dual problem $\mathcal{C}^* u = \mu u$ where $\mathcal{C}^* = \mathcal{F}^*\mathcal{B}^{-*}$ is the adjoint of $\mathcal{C}$. Note that $\mathcal{C}$ and $\mathcal{C}^*$ share the same spectral radius and μ° is also eigenvalue of $\mathcal{C}^*$ associated with a (different) positive eigenstate, denoted by u_*° vanishing only on Γ_+. Suppose

that $(\bar{\mu}, \bar{u})$ is any other eigenpair of $\mathcal{C}$ with $\bar{\mu} \neq \mu^\circ$ so that

$$\mu^\circ \langle \bar{u}, u_*^\circ \rangle = \langle \bar{u}, \mathcal{C}^* u_*^\circ \rangle = \langle \mathcal{C}\, \bar{u}, u_*^\circ \rangle = \bar{\mu} \langle \bar{u}, u_*^\circ \rangle$$

hence

$$(\bar{\mu}^{-1} - \mu^{\circ -1}) \langle \bar{u}, u_*^\circ \rangle = 0,$$

which, since $\bar{\mu} \neq \mu^\circ$, implies $\langle \bar{u}, u_*^\circ \rangle = 0$. This is impossible when $\bar{u}$ is non-trivial and non-negative, a contradiction.

It remains to prove that μ° is simple.

To that end, consider the operator $(\mathcal{F}\,\mathcal{B}^{-1})^* = \mathcal{B}^{-*}\,\mathcal{F}^*$ which has the same spectral radius as $\mathcal{F}\,\mathcal{B}^{-1}$ and as $\mathcal{B}^{-1}\,\mathcal{F}$, where the last fact follows from the general property that for bounded linear operators A, B the products AB and BA have the same spectral radius. This can be deduced by using that the spectral radius of an operator A equals $\lim_{n\to\infty} \|A^n\|^{1/n}$. Moreover, by the same arguments as before, $\mathcal{B}^{-*}\,\mathcal{F}^*$ has a positive eigenstate associated with the spectral radius μ° which we denote by $\hat{u^\circ}_*$.

Suppose now there exists another linearly independent eigenstate w° of $\mathcal{C}$, associated with μ°. Then we can orthonormalize w° in the eigenspace spanned by u° and w° yielding an eigenvector w in this subspace that satisfies $\langle u^\circ, w \rangle = 0$. Since u° is strictly positive (except on Γ_-) w must change sign. Since $|\,\mathcal{C}^2\, w| = |\,\mu^{\circ 2}\, w| = \mu^{\circ 2}\, |w|$, one observes that

$$E_1 = \langle |\,\mathcal{C}^2\, w|, \mathcal{F}^*\, \hat{u^\circ}_* \rangle = \mu^{\circ 2} \langle |w|, \mathcal{F}^*\, \hat{u^\circ}_* \rangle > 0. \tag{B2}$$

On the other hand,

$$E_2 = \langle \mathcal{C}^2\, |w|, \mathcal{F}^*\, \hat{u^\circ}_* \rangle = \langle |w|, (\mathcal{C}^*)^2\, \mathcal{F}^*\, \hat{u^\circ}_* \rangle = \langle |w|, \mathcal{F}^*\, \mathcal{B}^{-*}\, \mathcal{F}^*\, \mathcal{B}^{-*}\, \mathcal{F}^*\, \hat{u^\circ}_* \rangle. \tag{B3}$$

Moreover, note that $\mathcal{F}^*\, \hat{u^\circ}_*$ is a strictly positive eigenstate of $\mathcal{F}^*\, \mathcal{B}^{-*}$ because $\mathcal{F}^*\, \mathcal{B}^{-*}\, \mathcal{F}^*\, \hat{u^\circ}_* = \mathcal{F}^*(\mathcal{F}\,\mathcal{B}^{-1})^* \hat{u^\circ}_* = \mathcal{F}^*\, \mu^\circ\, \hat{u^\circ}_*$. Thus,

$$E_2 = \langle |w|, \mu^{\circ 2}\, \mathcal{F}^*\, \hat{u^\circ}_* \rangle = E_1 > 0.$$

Next one shows $E_1 < E_2$ which is the desired contradiction and confirms simplicity of μ°. To that end, since w changes sign, $|w| \pm v$ are both non-trivial non-negative functions. Note also that

$$\mathcal{B}^{-1} = (\mathcal{T}(\mathcal{I} - \mathcal{T}^{-1}\,\mathcal{K}))^{-1} = (\mathcal{I} - \mathcal{T}^{-1}\,\mathcal{K})^{-1}\,\mathcal{T}^{-1} = \sum_{j=0}^{\infty} (\mathcal{T}^{-1}\,\mathcal{K})^j\,\mathcal{T}^{-1} \geq \mathcal{T}^{-1},$$

where we have used Lemma 2.4. Now recall that $\mathcal{F}\,\mathcal{T}^{-1}\,\mathcal{F}$ is an integral operator with strictly positive kernel. Then

$$\begin{aligned}\mathcal{C}^2(|w| \pm w) &= (\mathcal{T}-\mathcal{K})^{-1}\,\mathcal{F}(\mathcal{T}-\mathcal{K})^{-1}\,\mathcal{F}(|w| \pm w)\\ &\geq (\mathcal{T}-\mathcal{K})^{-1}\,\mathcal{F}\,\mathcal{T}^{-1}\,\mathcal{F}(|w| \pm w) > 0 \quad \text{in } \overline{\mathrm{D}\times\mathrm{V}}\setminus\Gamma_-.\end{aligned}$$

This, in turn, implies

$$\mathcal{C}^2\,|w| > |\,\mathcal{C}^2\,w|,$$

and hence $E_2 > E_1$, the announced contradiction. □

Appendix C Proof of Remark 3.2

Regarding the interrelation between the inf-sup constant θ and the spectral gap Δ, stated in Remark 3.2, we first identify $\mathbb{U}_\circ^\perp$ in terms of Riesz projections. To that end, recall the two direct sum decompositions that

$$\mathbb{U} = \mathbb{U}_\circ \oplus_\perp \mathbb{U}_\circ^\perp = \mathbb{U}_\circ \oplus \mathbb{V}, \quad \mathbb{V} := \mathcal{E}_{\mathcal{C}}(\sigma(\mathcal{C})\setminus\{\mu^\circ\})\,\mathbb{U}\,. \tag{C1}$$

$\mathbb{U}_\circ^\perp$ and $\mathbb{V}$ both have co-dimension one but are generally different (they are equal when $\mathcal{C}$ is a normal operator). Observe next that

$$\mathbb{V}^* = \mathbb{U}_\circ^\perp, \quad \mathbb{V} = \langle u_*^\circ\rangle^\perp. \tag{C2}$$

To see this, we apply analogous properties of Riesz projections to the adjoint $\mathcal{C}^*$, where now $\mathbb{V}^*$ is the $\mathcal{C}^*$-invariant subspace associated with $\sigma(\mathcal{C}^*)\setminus\{u_*^\circ\}$, u_*° being the simple eigenvector of $\mathcal{C}^*$ associated with μ°. Then, for any $v \in \mathbb{V}$ and $\sigma_{>1} := \sigma(\mathcal{C})\setminus\{\mu^\circ\}$, one has

$$\begin{aligned}\langle u_*^\circ, v\rangle = \langle \mathcal{E}_{\mathcal{C}^*}(\mu^\circ)u_*^\circ, \mathcal{E}_{\mathcal{C}}(\sigma_{>1})v\rangle &= \langle u_*^\circ, \mathcal{E}^*_{\mathcal{C}^*}(\mu^\circ)\mathcal{E}_{\mathcal{C}}(\sigma_{>1})v\rangle\\ &= \langle u_*^\circ, \mathcal{E}_{\mathcal{C}}(\mu^\circ)\mathcal{E}_{\mathcal{C}}(\sigma_{>1})v\rangle = 0,\end{aligned}$$

where we have used (3.14) in the last step. The same reasoning applies to $\mathbb{U}_\circ^\perp = \langle u^\circ\rangle^\perp$, so that

$$u_*^\circ \perp \mathbb{V}, \quad u^\circ \perp \mathbb{V}^*, \tag{C3}$$

which is (C2).

The verification of the inequalities (3.26) will follow from judicious substitutes of the maximizer/minimizer in the inf-sup condition from these complement spaces. To that end, recall the meaning of $r_\circ \in \mathbb{U}_\circ^\perp = \mathbb{V}^*$, $u_\Lambda^* \in \mathbb{V}^* = \mathbb{U}_\circ^\perp$, λ_Λ from Remark 3.2. In particular, $\bar{\lambda}_\Lambda\,\mathcal{C}^*\,u_\Lambda^* = u_\Lambda^*$ so that u_Λ^* is also an eigenstate of $(\mathcal{M}_{\lambda^\circ}^\circ)^*$ with

eigenvalue $1-\lambda^\circ/\bar{\lambda}_\Lambda$. Then one deduces from (3.21) that

$$\theta = \sup_{\substack{v\in\mathbb{U}_\circ^\perp\\ \|v\|=1}} \langle \mathcal{M}_{\lambda^\circ} r_\circ, v\rangle \geq |\langle r_\circ, \mathcal{M}^*_{\lambda^\circ} u^*_\Lambda\rangle| = |1-\lambda^\circ/\bar{\lambda}_\Lambda||\langle r_\circ, u^*_\Lambda\rangle|$$
$$\geq (1-\lambda^\circ/|\bar{\lambda}_\Lambda|)\langle r_\circ, u^*_\Lambda\rangle|,$$

which is the lower estimate in (3.26).

To see the upper estimate in (3.26), denote by u_Λ an eigenvector of $\mathcal{C}$ associated with the largest in modulus eigenvalue $\mu_2 \in \sigma_{>1}$. Since u_2 belongs to $\mathbb{V}$, which generally differs from $\mathbb{U}_\circ^\perp = \mathbb{V}^*$, we let $z := P_{\mathbb{U}_\circ^\perp} u_\Lambda \in \mathbb{U}_\circ^\perp$, i.e., $u_\Lambda = z + P_{\mathbb{U}_\circ} u_\Lambda$. Since $\|z\| = \sqrt{\|u_\Lambda\|^2 - \|P_{\mathbb{U}_\circ} u_\Lambda\|^2} = \sqrt{1-\|P_{\mathbb{U}_\circ} u_\Lambda\|^2} =: s > 0$ and since

$$\mathcal{M}^\circ_{\lambda^\circ} z = \mathcal{M}^\circ_{\lambda^\circ}\left(u_\Lambda - P_{\mathbb{U}_\circ} u_\Lambda\right) = \mathcal{M}^\circ_{\lambda^\circ} u_\Lambda = (1-\lambda^\circ\mu_\Lambda)u_\Lambda$$

we conclude that

$$\theta \leq \sup_{\substack{v\in\mathbb{U}_\circ^\perp\\ \|v\|=1}} \frac{|\langle \mathcal{M}_{\lambda^\circ} z, v\rangle|}{s} = \sup_{\substack{v\in\mathbb{U}_\circ^\perp\\ \|v\|=1}} |1-\lambda^\circ\mu_2|s^{-1}|\langle u_\Lambda, v\rangle| = s^{-1}|1-\lambda^\circ\mu_2|\|z\|^2 = s\overline{\Delta},$$

which is the upper bound in (3.26).

When $\mathcal{C}$ is a normal operator $\mathbb{U}_\circ^\perp = \mathbb{U}_\circ^{*\perp}$ so that $\mathrm{dist}\big(\mathbb{U}_\circ^\perp, \mathbb{U}_\circ^{*\perp}\big) = 0$. Moreover, the smallest singular value of $\mathcal{M}^\circ_{\lambda^\circ}$ agrees then with its smallest eigenvalue which, in turn is one minus the largest eigenvalue of $\lambda^\circ \mathcal{C} u = \mu u$ which we know is $\lambda^\circ\lambda_\Lambda^{-1}$ with eigenvector u_Λ. Hence, in this case $\theta = \overline{\Delta} = \Delta$ which is (3.27).

The non-trivial gap in (3.26) can thus be viewed as quantifying deviation from normality. □

Acknowledgements Olga Mula kindly thanks the Smart State Chair from South Carolina University for funding her stay at that university in 2019 when the present research was initiated. We thank the reviewers for valuable suggestions that helped us to improve on the presentation of the material.

Funding This research was supported by the NSF Grants DMS 2038080, DMS-2012469, DMS-2245097, by the SmartState and Williams-Hedberg Foundation, by the SFB 1481, funded by the German Research Foundation. It was also funded by the Emergences Project Grant "Models and Measures" from the Paris City Council.

References

1. Allaire, G., Bal, G. (1999) Homogenization of the criticality spectral equation in neutron transport. ESAIM Mathematical Modelling and Numerical Analysis 33(4), 721–746
2. Avila, M., Codina, R., Principe, J.: Spatial approximation of the radiation transport equation using a subgrid-scale finite element method. Computer Methods in Applied Mechanics and Engineering 200(5-8), 425–438 (2011)
3. Agoshkov, A.: Boundary Value Problems for Transport Equations. Boston: Birkhäuser (1998)
4. Anselone, P., Rall, L.: The solution of characteristic value-vector problems by Newton's method. Numerische Mathematik 11, 38–45 (1968)
5. Ben-Artzi, J., Colbrook, M.J., Hansen, A.C., Nevanlinna, O., Seidel, M.: Computing spectra – on the solvability complexity index hierarchy and towers of algorithms. arXiv preprint arXiv:1508.03280 (2015)
6. Bal, G.: Couplage d'équations et homogénéisation en transport neutronique. PhD thesis, Paris VI (1997)
7. Bandtlow, O.F.: Estimates for norms of resolvents and an application to the perturbation of spectra. Mathematische Nachrichten 267(1), 3–11 (2004)
8. Bachmayr, M., Dahmen, W.: Adaptive Low-Rank Approximations for Operator Equations: Accuracy Control and Computational Complexity. Contemporary Mathematics **754**
9. Binev, P., Dahmen, W., DeVore, R.: Adaptive finite element methods with convergence rates. Numer. Math. 97, 219–268 (2004)
10. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods vol. 15. Springer (2012)
11. Badri, M.A., Jolivet, P., Rousseau, B., Favennec, Y.: High performance computation of radiative transfer equation using the finite element method. Journal of Computational Physics 360, 74–92 (2018)
12. Baudron, A.-M., Lautard, J.-J., Maday, Y., Mula, O.: MINARET: Towards a parallel 3D time-dependent neutron transport solver. In: SNA + MC 2013 - Joint International Conference on Supercomputing in Nuclear Applications + Monte Carlo, (2014)
13. Cohen, A., Dahmen, W., DeVore, R.: Adaptive wavelet methods II: Beyond the elliptic case. Foundations of Computational Mathematics 2, 203–245 (2002)
14. Cancès, E., Dusson, G., Maday, Y., Stamm, B., Vohralík, M.: Guaranteed and robust a posteriori bounds for laplace eigenvalues and eigenvectors: a unified framework. Numerische Mathematik 140(4), 1033–1079 (2018)
15. Dahmen, W., Gruber, F., Mula, O.: An adaptive nested source term iteration for radiative transfer equations. Mathematics of Computation 89(324), 1605–1646 (2020)
16. Dahmen, W., Harbrecht, H., Schneider, R.: Adaptive methods for boundary integral equations: Complexity and convergence estimates. Mathematics of Computation 76(259), 1243–1275 (2007)
17. Dahmen, W., Huang, C., Schwab, C., Welper, G.: Adaptive Petrov–Galerkin methods for first order transport equations. SIAM Journal on Numerical Analysis 50(5), 2420–2445 (2012)
18. Dahmen, W., Jürgens, M.: Error controlled regularization by projection. Electron. Trans. Numer. Anal. 25, 67–100 (2006)
19. Dautray, R., Lions, J.-L.: Mathematical Analysis and Numerical Methods for Science and Technology: Volume 6 Evolution Problems II. Springer (2012)
20. Dongarra, J.J., Moler, C.B., Wilkinson, J.H.: Improving the accuracy of computed eigenvalues and eigenvectors. SIAM Journal on Numerical Analysis 20(1), 23–45 (1983)
21. Dahmen, W., Reusken, A.: Numerik Für Ingenieure und Naturwissenschaftler, 2., korr. edn. Springer-Lehrbuch. Springer (2008)
22. Dunford, N., Schwartz, J.T.: Linear Operators, Part II.: Spectral Theory: Self Adjoint Operators in Hilbert Space vol. 10. Wiley-Interscience (1988)
23. Ern, A., Guermond, J.L.: Theory and Practice of Finite Elements vol. 159. Springer (2013)
24. Egger, H., Schlottbom, M.: A mixed variational framework for the radiative transfer equation. Math. Models Methods Appl. Sci. 22, 1150014–1150044 (2012)
25. Gillespie, T.A., Conway, J.B.: A course in functional analysis (graduate texts in mathematics 96, springer-verlag, 1985). Proceedings of the Edinburgh Mathematical Society 31(1), 166–166 (1988)
26. Giani, S., Graham, I.G.: A convergent adaptive method for elliptic eigenvalue problems. SIAM journal on numerical analysis 47(2), 1067–1091 (2009)
27. Gruber, F., Klewinghaus, A., Mula, O.: The DUNE-DPG library for solving PDEs with Discontinuous Petrov–Galerkin finite elements. Archive of Numerical Software 5(1), 111–128 (2017)

28. Golse, F., Lions, P.-L., Perthame, B., Sentis, R.: Regularity of the moments of the solution of a transport equation. Journal of functional analysis 76(1), 110–125 (1988)
29. Grella, K., Schwab, C.: Sparse discrete ordinates method in radiative transfer. Comp. Meth. in Applied Math. 11(3), 305–326 (2011)
30. Han, W.: A Posteriori Error Analysis Via Duality Theory: with Applications in Modeling and Numerical Approximations vol. 8. Springer (2004)
31. Han, W.: A posteriori error analysis in radiative transfer. Applicable Analysis 94(12), 2517–2534 (2015)
32. Jamelot, E., Dubois, J., Lautard, J.J., Calvin, C., Baudron, A.M.: High performance 3D neutron transport on peta scale and hybrid architectures within Apollo 3 code. In: International Conference on Mathematics and Computational Methods Applied to Nuclear Science and Engineering (2011)
33. Jürgens, M.: A semigroup approach to the numerical solution of parabolic differential equations. PhD thesis, Aachen (2005). Aachen, Techn. Hochsch., Diss., 2005
34. Jürgens, M.: Adaptive application of the operator exponential. Journal of Numerical Mathematics 14(3), 217–246 (2006)
35. Kanschat, G.: Solution of radiative transfer problems with finite elements. In: Numerical Methods in Multidimensional Radiative Transfer, pp. 49–98 (2009). Springer
36. Lewis, E.E., Miller, W.F.: Computational Methods of Neutron Transport. Wiley, New York (1984)
37. Madsen, N.K.: A posteriori error bounds for numerical solutions of the neutron transport equation. Mathematics of Computation **27**(124) (1973)
38. Mika, J.: Existence and uniqueness of the solution to the critical problem in the multigroup neutron-transport theory. Transport Theory and Statistical Physics 2(3), 243–270 (1972)
39. Mokhtar-Kharroubi, M.: Optimal spectral theory of the linear boltzmann equation. Journal of Functional Analysis 226(1), 21–47 (2005)
40. Manteuffel, T.A., Ressel, K.J., Starke, G.: A boundary functional for the least -squares finite-element solution of neutron transport problems. SIAM Journal on Numerical Analysis 37(2), 556–586 (1999)
41. Mula, O.: Some contributions towards the parallel simulation of time dependent neutron transport and the integration of observed data in real time. PhD thesis, Sorbonne University (2014)
42. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables. SIAM (2000)
43. Reuss, P.: Neutron Physics. EDP Sciences (2008)
44. Ragusa, J.C., Guermond, J.-L., Kanschat, G.: A robust S_N-DG-approximation for radiation transport in optically thick and diffusive regimes. Journal of Computational Physics 231(4), 1947–1962 (2012)
45. Scheben, F., Graham, I.G.: Iterative methods for neutron transport eigenvalue problems. SIAM Journal on Scientific Computing 33(5), 2785–2804 (2011)
46. Stacey, W.M.: Nuclear Reactor Physics. John Wiley & Sons (2018)
47. Stenger, F.: Numerical Methods Based on Sinc and Analytic Functions. Springer, New York (1993)
48. Tisseur, F.: Newton's method in floating point arithmetic and iterative refinement of generalized eigenvalue problems. SIAM Journal on Matrix Analysis and Applications 22(4), 1038–1057 (2001)
49. Wang, Y., Ragusa, J.C.: A high-order discontinuous galerkin method for the sn transport equations on 2d unstructured triangular meshes. Annals of Nuclear Energy 36(7), 931–939 (2009)