Hakula, Harri; Laaksonen, Mikael

Asymptotic convergence of spectral inverse iterations for stochastic eigenvalue problems

Published in:
Numerische Mathematik

DOI:
10.1007/s00211-019-01034-w

Published: 01/01/2019

Document Version
Publisher's PDF, also known as Version of record

Numerische Mathematik
https://doi.org/10.1007/s00211-019-01034-w

Numerische Mathematik

Asymptotic convergence of spectral inverse iterations for stochastic eigenvalue problems

Harri Hakula[1] · **Mikael Laaksonen**[1]

Received: 10 August 2017 / Revised: 9 January 2019

Abstract
We consider and analyze applying a spectral inverse iteration algorithm and its subspace iteration variant for computing eigenpairs of an elliptic operator with random coefficients. With these iterative algorithms the solution is sought from a finite dimensional space formed as the tensor product of the approximation space for the underlying stochastic function space, and the approximation space for the underlying spatial function space. Sparse polynomial approximation is employed to obtain the first one, while classical finite elements are employed to obtain the latter. An error analysis is presented for the asymptotic convergence of the spectral inverse iteration to the smallest eigenvalue and the associated eigenvector of the problem. A series of detailed numerical experiments supports the conclusions of this analysis.

Mathematics Subject Classification 65C20 · 65N12 · 65N15 · 65N25 · 65N30

1 Introduction

During the recent years numerical solution of stochastic partial differential equations (sPDE) has attracted a lot of attention and become a well-established field. However, the field of stochastic eigenvalue problems (sEVP) and their numerical solution is still in its infancy. It is natural that, after the source problem, more effort is put on addressing the eigenvalue problem.

Harri Hakula: The work of this author was supported by the (FP7/2007–2013) ERC Grant Agreement No. 339380. Mikael Laaksonen: The work of this author was supported by the Magnus Ehrnrooth Foundation.

✉ Harri Hakula
harri.hakula@aalto.fi

Mikael Laaksonen
mikael.j.laaksonen@aalto.fi

[1] Department of Mathematics and Systems Analysis, Aalto University, Espoo, Finland

Published online: 19 March 2019

A few different algorithms have recently been suggested for computing approximate eigenpairs of sEVPs. As with sPDEs, the solution methods are typically divided into intrusive and non-intrusive ones. A benchmark for non-intrusive methods is the sparse collocation algorithm suggested and thoroughly analyzed by Andreev and Schwab [1]. An attempt towards a Galerkin-based (intrusive) method was made by Verhoosel et al. [20], though this method omits uniform normalization of the eigenmodes. Very recently Meidani and Ghanem proposed a spectral power iteration, in which the eigenmodes are normalized using a quadrature rule over the parameter space [16]. The algorithm has been further developed and studied by Sousedík and Elman [19]. However, neither of the papers present a comprehensive error analysis for the method.

Inspired by the original method of Meidani and Ghanem we have suggested a purely Galerkin-based spectral inverse iteration, in which normalization of the eigenmodes is achieved via solution of a simple nonlinear system [11]. This method, and its generalization to a spectral subspace iteration, is the focus of the current paper. Although the algorithms in [16,19] differ from ours in the way normalization is performed, the basic principles are still the same and hence our results on convergence should apply to these methods as well.

In this work we consider computing eigenpairs of an elliptic operator with random coefficients. We assume a physical domain $D \subset \mathbb{R}^d$ and, in order to capture the random dimension of the system, a parameter domain $\Gamma \subset \mathbb{R}^\infty$ with associated measure ν. One may think of a parametrization that arises from Karhunen-Loève representations of the random coefficients in the system, for instance. Discretization in space is achieved by standard FEM and associated with a discretization parameter h, whereas discretization in the random dimension is achieved using collections of certain multivariate polynomials. These collections are represented by multi-index sets $\mathcal{A}_\epsilon$ of increasing cardinality $\#\mathcal{A}_\epsilon$ as $\epsilon \to 0$.

In the current paper we present a step-by-step analysis that leads to the main result: the asymptotic convergence of the spectral inverse iteration towards the exact eigenpair (μ, u). In this context the eigenpair of interest is the ground state, i.e., the smallest eigenvalue and the associated eigenfunction of the system. However, analogously to the classical inverse iteration, the computation of other eigenpairs may be possible by using a suitably chosen shift parameter $\lambda \in \mathbb{R}$. We show that under sufficient assumptions the iterates of the algorithm (μ_k, u_k) for $k = 1, 2, \ldots$ obey

$$||u - u_{h,\mathcal{A},k}||_{L^2_\nu(\Gamma)\otimes L^2(D)} \lesssim h^{1+l} + (\#\mathcal{A}_\epsilon)^{-r} + \lambda_{1/2}^k \tag{1}$$

and

$$||\mu - \mu_{h,\mathcal{A},k}||_{L^2_\nu(\Gamma)} \lesssim h^{2l} + (\#\mathcal{A}_\epsilon)^{-r} + \lambda_{1/2}^k, \tag{2}$$

where $l \in \mathbb{N}$ is the degree of polynomials used in the spatial discretization and $r > 0$ depends on the properties of the region to which the solution, as a function of the parameter vector, admits a complex-analytic extension. The quantity $\lambda_{1/2}$ reflects the gap between the two smallest eigenvalues of the system and should be less than one.

The first term in the formulas (1) and (2) is justified by standard theory for Galerkin approximation of eigenvalue problems, a simple consequence of which we have recapped in Theorem 1. The second term can be deduced from Theorem 2, which

bounds the Galerkin approximation errors by residuals of certain polynomial approximations of the solution. Using best P-term polynomial approximations, we see that these residuals are ultimately expected to decay at an algebraic rate $r > 0$, see [5] and [7]. Finally, the third term follows from Theorem 3, which states that asymptotically the iterates of the spectral inverse iteration converge to a fixed point in geometric fashion. Here the analogy to classical inverse iteration is evident. Each of these three important steps that comprise the main result is separately verified through detailed numerical examples.

A variant of our algorithm for spectral subspace iteration is also presented. No analysis of this algorithm is given, but the numerical experiments support the conclusion that it converges towards the exact subspace of interest, and that the rate of convergence is analogous to what we would expect from classical theory. This is despite the fact that the individual eigenmodes, as defined by the pointwise order of magnitude of the eigenvalues, are not continuous functions over the parameter space due to an eigenvalue crossing. To the authors' knowledge such a scenario has not yet been considered in the scientific literature.

The rest of the paper is organized as follows. Our model problem and its fundamental properties are assessed in Sects. 2 and 3. A detailed review of the discretization of the spatial and stochastic approximation spaces is given in Sect. 4. Analysis of the spectral inverse iteration, supported by thorough numerical experiments, is given in Sect. 5. Finally, the algorithm of spectral subspace iteration and numerical experiments of its convergence are presented in Sect. 6.

2 Problem statement

In this work we consider eigenvalue problems of elliptic operators with random coefficients. It is assumed that the random coefficients admit a parametrization with respect to countably many independent and bounded random variables. As a model problem we consider the eigenvalue problem of a diffusion operator with a random diffusion coefficient. It will be evident, however, that our methods and analysis in fact cover a much broader class of problems.

2.1 Model problem

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space, Ω being the set of outcomes, $\mathcal{F}$ a σ-algebra of events, and $\mathcal{P}$ a probability measure defined on Ω. We denote by $L^2_{\mathcal{P}}(\Omega)$ the space of square integrable random variables on Ω and define for $v \in L^2_{\mathcal{P}}(\Omega)$ the expected value

$$\mathbb{E}[v] = \int_{\Omega} v(\omega)\, d\mathcal{P}(w)$$

and variance $\mathrm{Var}[v] = \mathbb{E}[(v - \mathbb{E}[v])^2]$.

Let $D \subset \mathbb{R}^d$ be a bounded convex domain with a sufficiently smooth boundary and assume a diffusion coefficient $a : D \times \Omega \to \mathbb{R}$ that is a random field on D.

The diffusion coefficient is assumed to be strictly uniformly positive and uniformly bounded, i.e., for some positive constants $a_{\min}$ and $a_{\max}$ it holds that

$$\mathcal{P}\left(\omega \in \Omega : a_{\min} \leq \operatorname*{ess\,inf}_{x\in D} a(x,\omega) \leq \operatorname*{ess\,sup}_{x\in D} a(x,\omega) \leq a_{\max}\right) = 1. \tag{3}$$

We now formulate the model problem as: find functions $\mu : \Omega \to \mathbb{R}$ and $u : D \times \Omega \to \mathbb{R}$ such that the equations

$$\begin{cases} -\nabla \cdot (a(\cdot,\omega)\nabla u(\cdot,\omega)) = \mu(\omega) u(\cdot,\omega) & \text{in } D \\ u(\cdot,\omega) = 0 & \text{on } \partial D \end{cases} \tag{4}$$

hold $\mathcal{P}$-almost surely. In order to make the solutions physically meaningful we also impose a normalization condition $||u(\cdot,\omega)||_{L^2(D)} = 1$ that should hold $\mathcal{P}$-almost surely.

2.2 Parametrization of the random input

We make the assumption that the input random field admits a representation of the form

$$a(x,\omega) = a_0(x) + \sum_{m=1}^{\infty} a_m(x) y_m(\omega), \tag{5}$$

where $\{y_m\}_{m=1}^{\infty}$ are mutually independent and bounded random variables. For simplicity, we assume here that each y_m is uniformly distributed. Thus, after possible rescaling, the dependence on ω is now parametrized by the vector $y = (y_1, y_2, \ldots) \in \Gamma := [-1,1]^{\infty}$. We denote by ν the underlying uniform product probability measure and by $L^2_\nu(\Gamma)$ the corresponding weighted L^2-space.

The usual convention is that the parametrization (5) results from a Karhunen-Loève expansion, which gives $a(x,\omega)$ as a linear combination of the eigenfunctions of the associated covariance operator. The distinguishing feature of the Karhunen-Loève expansion compared to other linear expansions is that it minimizes the mean square truncation error [9].

It is easy to see that $a_0 \in L^\infty(D)$ and

$$\operatorname*{ess\,inf}_{x\in D} a_0(x) > \sum_{m=1}^{\infty} ||a_m||_{L^\infty(D)} \tag{6}$$

are sufficient conditions to ensure the assumption (3). In order to ensure analyticity of the eigenpair (μ, u) with respect to the parameter vector $y = (y_1, y_2, \ldots)$ we assume that

$$\sum_{m=1}^{\infty} ||a_m||^{p_0}_{L^\infty(D)} < \infty \tag{7}$$

for some $p_0 \in (0, 1)$ and that for a certain level of smoothness $s \in \mathbb{N}$ we have $a_0 \in W^{s,\infty}(D)$ and

$$\sum_{m=1}^{\infty} ||a_m||^{p_s}_{W^{s,\infty}(D)} < \infty \tag{8}$$

for some $p_s \in (0, 1)$. In particular, we consider the interesting case of algebraic

$$||a_m||_{L^\infty(D)} \leq Cm^{-\varsigma}, \quad \varsigma > 1, \quad m = 1, 2, \ldots$$

decay of the coefficients in the series (5).

2.3 Parametric eigenvalue problem and its variational formulation

With the diffusion coefficient given by (5), the model problem (4) becomes an eigenvalue problem of the operator

$$(A(y)v)(x) := -\nabla \cdot (a(x, y)\nabla v(x)), \quad x \in D, \quad y \in \Gamma,$$

where

$$a(x, y) = a_0(x) + \sum_{m=1}^{\infty} a_m(x) y_m.$$

Thus, we obtain the parametric eigenvalue problem: find $\mu : \Gamma \to \mathbb{R}$ and $u : \Gamma \to H_0^1(D)$ such that

$$A(y)u(y) = \mu(y)u(y) \quad \forall y \in \Gamma. \tag{9}$$

We denote by $\sigma(A(y))$ the set of eigenvalues of $A(y)$ for $y \in \Gamma$.

For any fixed $y \in \Gamma$ the problem (9) reduces to a single deterministic eigenvalue problem. In variational form this is given by: find $\mu(y) \in \mathbb{R}$ and $u(\cdot, y) \in H_0^1(D)$ such that

$$b(y; u(\cdot, y), v) = \mu(y)\langle u(\cdot, y), v\rangle_{L^2(D)} \quad \forall v \in H_0^1(D), \tag{10}$$

where

$$b(y; u, v) := \int_D a(\cdot, y)\nabla u \cdot \nabla v \, dx.$$

Under assumption (6) the bilinear form $b(y; u, v)$ is continuous and elliptic. Thus, as in [1,11], we deduce that the problem (10) admits a countable number of real eigenvalues and corresponding eigenfunctions that form an orthogonal basis of $L^2(D)$.

3 Analyticity of eigenmodes

A key issue in the analysis of parametric eigenvalue problems is that eigenvalues may cross within the parameter space. Here we first disregard this possibility and recap

the main results from [1] for simple eigenvalues that are sufficiently well separated from the rest of the spectrum. In Sect. 6 we briefly comment on the case of possibly clustered eigenvalues and associated invariant subspaces.

We call an eigenvalue μ of problem (9) strictly nondegenerate if

(i) $\mu(y)$ is simple as an eigenvalue of $A(y)$ for all $y \in \Gamma$ and
(ii) the minimum spectral gap $\inf_{y\in\Gamma} \operatorname{dist}(\mu(y), \sigma(A(y))\backslash\{\mu(y)\})$ is positive.

In the case of strictly nondegenerate eigenvalues, the eigenpair (μ, u) is in fact analytic with respect to the parameter vector y.

Proposition 1 *Consider a strictly nondegenerate eigenvalue μ of the problem (9) and the corresponding eigenfunction u normalized so that $||u(y)||_{L^2(D)} = 1$ for all $y \in \Gamma$. For $s \in \mathbb{N}$ assume that $a_0 \in W^{s,\infty}(D)$ and the assumptions (6)–(8) hold for some $p_0, p_s \in (0, 1)$. Given $\tau = (\tau_1, \tau_2, \ldots) \in \mathbb{R}_+^\infty$ define*

$$E(\tau) := \{z \in \mathbb{C}^\infty \mid \operatorname{dist}(z_m, [-1, 1]) \leq \tau_m\}.$$

Then there exists $C_1 > 0$ independent of m such that with $C_2 > 0$ arbitrary and τ given by

$$\tau_m = \min\left\{C_1||a_m||_{L^\infty(D)}^{p_0-1}, C_2||a_m||_{W^{s,\infty}(D)}^{p_s-1}\right\}, \quad m = 1, 2, \ldots$$

the eigenpair (μ, u) can be extended to a jointly complex-analytic function on $E(\tau)$ with values in $\mathbb{C} \times (H^{s+1}(D) \cap H_0^1(D))$.

Proof This is analogous to Corollary 2 of Theorem 4 in [1]. □

It is well known that for elliptic operators on a connected domain D the smallest eigenvalue is simple [12]. Thus, Proposition 1 may at least be applied for the smallest eigenvalue of problem (9).

4 Stochastic finite elements

Proposition 1, under sufficient assumptions, guarantees the existence of an analytic eigenpair for problem (9). It now makes sense to look for the eigenvalue in the space $L_\nu^2(\Gamma)$ and the eigenfunction in the space $L_\nu^2(\Gamma) \otimes H_0^1(D)$. The space $H_0^1(D)$ may be discretized by means of the traditional finite element method. For the discretization of $L_\nu^2(\Gamma)$, we follow the usual convention in stochastic Galerkin methods and construct a basis of orthogonal polynomials of the input random variables. Orthogonal polynomials for various probability distributions exist and the use of these as the approximation basis has been observed to yield optimal rates of convergence [18,21]. Here we consider uniformly distributed random variables which lead to the choice of tensorized Legendre polynomials.

4.1 Galerkin discretization in space

Let $V_h \subset H_0^1(D)$ denote a finite dimensional approximation space associated with the discretization parameter $h > 0$. We assume approximation estimates

$$\inf_{v_h \in V_h} ||v - v_h||_{L^2(D)} \leq Ch^{l+1}||v||_{H^{l+1}(D)} \tag{11}$$

and

$$\inf_{v_h \in V_h} ||v - v_h||_{H_0^1(D)} \leq Ch^{l}||v||_{H^{l+1}(D)} \tag{12}$$

that are standard for piecewise polynomials of degree l.

Fix $y \in \Gamma$ and let (μ_h, u_h) be the solution to the variational equation

$$b(y; u_h(\cdot, y), v_h) = \mu_h(y)\langle u_h(\cdot, y), v_h\rangle_{L^2(D)} \quad \forall v_h \in V_h, \tag{13}$$

where $b(y; \cdot, \cdot)$ is as in (10). Then we have the following bounds for the discretization error.

Theorem 1 *Assume* (11) *and* (12). *For* $y \in \Gamma$ *let* $\mu(y)$ *be a simple eigenvalue of* (10) *and* $\mu_h(y)$ *an eigenvalue of* (13) *such that* $\lim_{h\to 0} \mu_h(y) = \mu(y)$. *Let* $u(\cdot, y) \in H^{1+l}(D)$ *and* $u_h(\cdot, y) \in V_h$ *denote the associated eigenfunctions normalized in* $L^2(D)$. *Then there exists* $C > 0$ *such that*

$$|\mu(y) - \mu_h(y)| \leq Ch^{2l}, \tag{14}$$

and

$$||u(\cdot, y) - u_h(\cdot, y)||_{L^2(D)} \leq Ch^{1+l}||u(\cdot, y)||_{H^{1+l}(D)}. \tag{15}$$

as $h \to 0$.

Proof This follows from the theory of Galerkin approximation for variational eigenvalue problems. See Section 8 in [3] and Section 9 in [8]. □

Let $V_h = \text{span}\{\varphi_i\}_{i\in J}$ where $J := \{1, 2, \ldots, N\}$. Then (13) can be written as a parametric matrix eigenvalue problem: find $\mu_h : \Gamma \to \mathbb{R}$ and $\mathbf{u}_h : \Gamma \to \mathbb{R}^N$ such that

$$\left(\mathbf{K}^{(0)} + \sum_{m=1}^{\infty} \mathbf{K}^{(m)} y_m\right)\mathbf{u}_h(y) = \mu_h(y)\mathbf{M}\mathbf{u}_h(y) \quad \forall y \in \Gamma, \tag{16}$$

where $u_h(x, y) = \sum_{i\in J} \varphi_i(x)(\mathbf{u}_h)_i(y)$. The coefficient matrices are given by

$$\mathbf{K}_{ij}^{(m)} = \int_D a_m \nabla\varphi_i \cdot \nabla\varphi_j \, dx, \quad m = 0, 1, \ldots$$

and

$$\mathbf{M}_{ij} = \int_D \varphi_i\varphi_j \, dx.$$

For each fixed $y \in \Gamma$ the problem (16) reduces to a positive-definite generalized matrix eigenvalue problem.

4.2 Legendre chaos

Recall that $y = (y_1, y_2, \ldots) \in \Gamma$ is a vector of mutually independent uniform random variables and ν is the underlying constant product probablity measure. Now

$$\mathbb{E}[v] = \int_\Gamma v(y)\, d\nu(y) \tag{17}$$

whenever the integral is finite. We define $(\mathbb{N}_0^\infty)_c$ to be the set of all multi-indices with finite support, i.e.,

$$(\mathbb{N}_0^\infty)_c := \{\alpha \in \mathbb{N}_0^\infty \mid \#\,\mathrm{supp}(\alpha) < \infty\},$$

where $\mathrm{supp}(\alpha) = \{m \in \mathbb{N} \mid \alpha_m \neq 0\}$. Given a multi-index $\alpha \in (\mathbb{N}_0^\infty)_c$ we now define the multivariate Legendre polynomial

$$\Lambda_\alpha(y) := \prod_{m \in \mathrm{supp}\,\alpha} L_{\alpha_m}(y_m),$$

where $L_p(x)$ denotes the univariate Legendre polynomial of degree p. We will assume the normalization $\mathbb{E}[\Lambda_\alpha^2] = 1$ for all $\alpha \in (\mathbb{N}_0^\infty)_c$.

The system $\{\Lambda_\alpha(y) \mid \alpha \in (\mathbb{N}_0^\infty)_c\}$ forms an orthonormal basis of $L_\nu^2(\Gamma)$. Therefore, we may write any square integrable random variable v in a series

$$v(y) = \sum_{\alpha \in (\mathbb{N}_0^\infty)_c} v_\alpha \Lambda_\alpha(y) \tag{18}$$

with convergence in $L_\nu^2(\Gamma)$. The expansion coefficients are given by $v_\alpha = \mathbb{E}[v\Lambda_\alpha]$.

Due to the orthogonality of the Legendre polynomials we have $\mathbb{E}[\Lambda_\alpha] = \delta_{\alpha 0}$ and $\mathbb{E}[\Lambda_\alpha \Lambda_\beta] = \delta_{\alpha\beta}$ for all $\alpha, \beta \in (\mathbb{N}_0^\infty)_c$. Moreover, we denote

$$\begin{aligned} c_{\alpha\beta\gamma} &:= \mathbb{E}[\Lambda_\alpha \Lambda_\beta \Lambda_\gamma], \quad \alpha, \beta, \gamma \in (\mathbb{N}_0^\infty)_c \\ c_{m\alpha\beta} &:= \mathbb{E}[y_m \Lambda_\alpha \Lambda_\beta], \quad m \in \mathbb{N}, \quad \alpha, \beta \in (\mathbb{N}_0^\infty)_c \\ c_{0\alpha\beta} &:= \delta_{\alpha\beta}, \quad \alpha, \beta \in (\mathbb{N}_0^\infty)_c. \end{aligned}$$

4.3 Sparse polynomial approximation in the parameter domain

We fix a finite set $\mathcal{A} \subset (\mathbb{N}_0^\infty)_c$ and employ the approximation space $W_\mathcal{A} = \mathrm{span}\{\Lambda_\alpha\}_{\alpha \in \mathcal{A}} \subset L_\nu^2(\Gamma)$. We let $P_\mathcal{A}$ and $R_\mathcal{A}$ denote the underlying projection and residual operators so that $v \in L_\nu^2(\Gamma)$ is approximated by

$$P_{\mathcal{A}}(v)(y) = \sum_{\alpha \in \mathcal{A}} v_\alpha \Lambda_\alpha(y)$$

and the approximation error is given by $R_{\mathcal{A}}(v) = v - P_{\mathcal{A}}(v)$. Since

$$||R_{\mathcal{A}}(v)||^2_{L^2_\nu(\Gamma)} = \mathbb{E}\left[\left(\sum_{\alpha \in \mathcal{A}^c} v_\alpha \Lambda_\alpha\right)^2\right] = \sum_{\alpha \in \mathcal{A}^c} v_\alpha^2, \tag{19}$$

where $\mathcal{A}^c = \{\alpha \in (\mathbb{N}_0^\infty)_c \mid \alpha \notin \mathcal{A}\}$, we conclude that the choice of the multi-index set $\mathcal{A}$ ultimately determines the accuracy of our expansion.

We proceed as in [5] and use best P-term approximations to prove convergence of the approximation error.

Proposition 2 *Let H be a Hilbert space. Assume that $v\colon \Gamma \to H$ admits a complex-analytic extension in the region*

$$E(\tau) := \{z \in \mathbb{C}^\infty \mid \operatorname{dist}(z_m, [-1, 1]) \leq \tau_m\}$$

with

$$\tau_m \geq C m^\varrho, \quad \varrho > 1, \quad m = 1, 2, \ldots \tag{20}$$

Given $\epsilon > 0$ define

$$\mathcal{A}_\epsilon := \left\{ \alpha \in (\mathbb{N}_0^\infty)_c \;\middle|\; \prod_{m \in \operatorname{supp} \alpha} \eta_m^{\alpha_m} > \epsilon \right\},$$

where

$$\eta_m := \left(\tau_m + \sqrt{1 + \tau_m^2}\right)^{-1}, \quad m = 1, 2, \ldots$$

Then

$$||R_{\mathcal{A}_\epsilon}(v)||_{L^2_\nu(\Gamma)\otimes H} \leq \epsilon ||v||_{L^\infty(E(\tau);H)} \tag{21}$$

and as $\epsilon \to 0$ we have

$$\#\mathcal{A}_\epsilon \leq C(\varrho, r)\epsilon^{-1/r} \tag{22}$$

for any $0 < r < \varrho - \frac{1}{2}$.

Proof Fix $\epsilon > 0$ and let $P = \#\mathcal{A}_\epsilon$. Set $M = \max\{m \in \mathbb{N} \mid \exists \alpha \in \mathcal{A}_\epsilon \text{ s.t. } \alpha_m \neq 0\}$ and $v_M(z) = v(z_1, \ldots, z_M, 0, 0, \ldots)$ so that $P_{\mathcal{A}_\epsilon}(v) = P_{\mathcal{A}_\epsilon}(v_M)$. The norm of the residual may now be separated into two parts in the following sense

$$||R_{\mathcal{A}_\epsilon}(v)||_{L^2_\nu(\Gamma)\otimes H} \leq ||v - v_M||_{L^2_\nu(\Gamma)\otimes H} + ||v_M - P_{\mathcal{A}_\epsilon}(v_M)||_{L^2_\nu(\Gamma)\otimes H}. \tag{23}$$

For the second term we may apply the proof of Proposition 3.1 in [6] and obtain

$$||v_M - P_{\mathcal{A}_\epsilon}(v_M)||_{L^2_\nu(\Gamma)\otimes H} \leq C(\varrho, r)P^{-r}||v_M||_{L^\infty(E(\tau);H)}. \tag{24}$$

On the other hand, in order to bound the first term we note that

$$\sum_{m>M}^{\infty}(\eta_m - 1)^{-1} \leq C\sum_{m>M}^{\infty} m^{-\varrho} \leq C\int_M^\infty x^{-\varrho}\,dx \leq C(\varrho)M^{1-\varrho}. \tag{25}$$

Thus, by Lemmas 4.3. and 4.4 in [2], we obtain

$$||v - v_M||_{L^2_\nu(\Gamma)\otimes H} \leq C||v||_{L^\infty(E(\tau);H)}\sum_{m>M}^{\infty}(\eta_m - 1)^{-1} \leq C(\varrho)P^{-r}||v||_{L^\infty(E(\tau);H)} \tag{26}$$

for any $M \geq C(\varrho)P^{r/(\varrho-1)}$. The claim follows from combining (24) and (26). □

4.4 Stochastic Galerkin approximation of vectors and matrices

We now generalize the concept of sparse polynomial approximation to vector and matrix valued functions. Assume that the dimensions of the approximation spaces V_h and $W_\mathcal{A}$ are N and P respectively. We denote by $W_\mathcal{A}^N$ (or $W_\mathcal{A}^{N\times N}$) the space of functions $\mathbf{v} : \Gamma \to \mathbb{R}^N$ (or $\mathbf{A} : \Gamma \to \mathbb{R}^{N\times N}$) whose every component is in $W_\mathcal{A}$. Whenever $\mathbf{v} \in W_\mathcal{A}^N$ and $\alpha \in \mathcal{A}$ we set $v_{\alpha i} = (\mathbf{v}_i)_\alpha$ and use $\mathbf{v}_\alpha$ to denote the vector of coefficients $\{v_{\alpha i}\}_{i\in J} \in \mathbb{R}^N$. Moreover, we associate any $v \in W_\mathcal{A}$ with the array of coefficients $\hat{v} := \{v_\alpha\}_{\alpha\in\mathcal{A}} \in \mathbb{R}^P$ and similarly any $\mathbf{v} \in W_\mathcal{A}^N$ with the array of coefficients $\hat{\mathbf{v}} := \{v_{\alpha i}\}_{\alpha\in\mathcal{A}, i\in J} \in \mathbb{R}^{PN}$.

We denote by $\langle\cdot,\cdot\rangle_{\mathbb{R}^N_\mathbf{M}}$ the inner product on $\mathbb{R}^N$ induced by the positive definite matrix $\mathbf{M}$ and by $||\cdot||_{\mathbb{R}^N_\mathbf{M}}$ the associated norm. Furthermore, we let $||\cdot||_{\mathbb{R}^P}$ denote the standard norm on $\mathbb{R}^P$ and $||\cdot||_{\mathbb{R}^P\otimes\mathbb{R}^N_\mathbf{M}}$ denote the tensorized norm on $\mathbb{R}^{PN}$ given by

$$||\hat{\mathbf{v}}||^2_{\mathbb{R}^P\otimes\mathbb{R}^N_\mathbf{M}} := \sum_{\alpha\in\mathcal{A}}\sum_{i\in J}\sum_{j\in J} v_{\alpha i}\mathbf{M}_{ij}v_{\alpha j}.$$

Remark 1 Observe that if $v \in W_\mathcal{A}\otimes V_h$ is written as $v(x,y) = \sum_{i\in J}\varphi_i(x)\mathbf{v}_i(y)$, then

$$||v||^2_{L^2_\nu(\Gamma)\otimes L^2(D)} = ||\mathbf{v}||^2_{L^2_\nu(\Gamma)\otimes\mathbb{R}^N_\mathbf{M}} = ||\hat{\mathbf{v}}||^2_{\mathbb{R}^P\otimes\mathbb{R}^N_\mathbf{M}}. \tag{27}$$

Let us consider the linear system defined by a parametric matrix $\mathbf{A} \in W_\mathcal{A}^{N\times N}$. The Galerkin approximation of this system is: given $\mathbf{f} \in W_\mathcal{A}^N$ find $\mathbf{v} \in W_\mathcal{A}^N$ such that

$$P_\mathcal{A}(\mathbf{A}\mathbf{v})(y) = \mathbf{f}(y) \quad \forall y \in \Gamma. \tag{28}$$

We define moment matrices $G^{(m)} \in \mathbb{R}^{P\times P}$ for $m \in \mathbb{N}_0$ and $G^{(\alpha)} \in \mathbb{R}^{P\times P}$ for $\alpha \in \mathcal{A}$ by setting $[G^{(m)}]_{\alpha\beta} = c_{m\alpha\beta}$ and $[G^{(\alpha)}]_{\beta\gamma} = c_{\alpha\beta\gamma}$. Using this notation we may write (28) as the fully discrete system: given $\hat{\mathbf{f}} \in \mathbb{R}^{PN}$ find $\hat{\mathbf{v}} \in \mathbb{R}^{PN}$ such that

$$\left(\sum_{\alpha\in\mathcal{A}} G^{(\alpha)}\otimes \mathbf{A}_\alpha\right)\hat{\mathbf{v}} = \hat{\mathbf{f}}, \tag{29}$$

where $\mathbf{A}_\alpha = \mathbb{E}[\mathbf{A}\Lambda_\alpha] \in \mathbb{R}^{N\times N}$. The existence of a solution, i.e. the invertibility of the coefficient matrix, is guaranteed by the following lemma.

Lemma 1 *If $\mathbf{A} \in W_{\mathcal{A}}^{N\times N}$ is a parametric matrix such that $\mathbf{A}(y)$ is positive-definite for every $y \in \Gamma$, then for any $\mathbf{f} \in W_{\mathcal{A}}^{N}$ there exists a unique $\mathbf{v} \in W_{\mathcal{A}}^{N}$ such that (28) holds. Furthermore,*

$$||\mathbf{v}||_{L^2_\nu(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} \leq \sup_{y\in\Gamma} \lambda^{-1}(y)||\mathbf{f}||_{L^2_\nu(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}}, \tag{30}$$

where $\lambda(y)$ is the smallest eigenvalue of $\mathbf{A}(y)$ for each $y \in \Gamma$.

Proof Observe that the system (28) is equivalent to the variational form

$$\mathbb{E}[\mathbf{v}^T\mathbf{A}\mathbf{w}] = \mathbb{E}[\mathbf{f}^T\mathbf{w}] \quad \forall \mathbf{w} \in W_{\mathcal{A}}^{N}. \tag{31}$$

The left hand side of (31) is a symmetric and elliptic bilinear form so the existence of a unique solution is guaranteed by the Lax-Milgram Lemma. Moreover, the associated coefficient matrix in (29) is positive definite.

Now let $\tilde{\lambda} \in \mathbb{R}$ be such that $\tilde{\lambda} < \inf_{y\in\Gamma} \lambda(y)$. The matrix $\mathbf{A}(y) - \tilde{\lambda}\mathbf{I}_N$, where $\mathbf{I}_N$ is the identity matrix, is positive definite for all $y \in \Gamma$. Thereby the eigenvalues of the associated coefficient matrix should be positive. Let χ be an eigenvalue of (29), i.e., there exists $\mathbf{w} \in W_{\mathcal{A}}^{N}$ such that

$$P_{\mathcal{A}}(\mathbf{A}\mathbf{w})(y) = \chi\,\mathbf{w}(y) \quad \forall y \in \Gamma. \tag{32}$$

Then

$$P_{\mathcal{A}}((\mathbf{A} - \tilde{\lambda}\mathbf{I}_N)\mathbf{w})(y) = P_{\mathcal{A}}(\mathbf{A}\mathbf{w})(y) - \tilde{\lambda}\mathbf{w}(y) = (\chi - \tilde{\lambda})\mathbf{w}(y) \quad \forall y \in \Gamma \tag{33}$$

and we deduce that $\chi > \tilde{\lambda}$. Equation (30) now follows from taking the limit $\tilde{\lambda} \to \inf_{y\in\Gamma} \lambda(y)$. □

5 Spectral inverse iteration

In this section we introduce the algorithm of spectral inverse iteration, analyze its asymptotic convergence, and present numerical examples to support our analysis. The spectral inverse iteration, see [11], can be considered as an extension of the classical inverse iteration to the case of parametric matrix eigenvalue problems. In the spectral version each of the elementary operations is computed in Galerkin sense via projecting to the sparse polynomial basis $W_{\mathcal{A}}$. Optimal convergence of the algorithm requires that the eigenmode of interest, i.e., the smallest eigenvalue of the parametric matrix, is strictly nondegenerate.

5.1 Algorithm description

Fix a finite set of multi-indices $\mathcal{A} \subset (\mathbb{N}_0^\infty)_c$ and let $P = \#\mathcal{A}$. The spectral inverse iteration for the system (16) is now defined in Algorithm 1. One should note that, if the projections in the algorithm were precise, the algorithm would correspond to performing classical inverse iteration pointwise over the parameter space Γ. We expect the algorithm to converge to an approximation of the eigenvector corresponding to the smallest eigenvalue of the system.

Algorithm 1 (Spectral inverse iteration) *Fix* $tol > 0$ *and let* $\mathbf{u}^{(0)} \in W_{\mathcal{A}}^N$ *be an initial guess for the eigenvector. For* $k = 1, 2, \ldots$ *do*

(1) *Solve* $\mathbf{v} \in W_{\mathcal{A}}^N$ *from the linear equation*

$$P_{\mathcal{A}}\left(\mathbf{K}\mathbf{v}\right) = \mathbf{M}\mathbf{u}^{(k-1)}. \tag{34}$$

(2) *Solve* $s \in W_{\mathcal{A}}$ *from the nonlinear equation*

$$P_{\mathcal{A}}(s^2) = P_{\mathcal{A}}\left(||\mathbf{v}||^2_{\mathbb{R}^N_{\mathbf{M}}}\right). \tag{35}$$

(3) *Solve* $\mathbf{u}^{(k)} \in W_{\mathcal{A}}^N$ *from the linear equation*

$$P_{\mathcal{A}}\left(s\mathbf{u}^{(k)}\right) = \mathbf{v}. \tag{36}$$

(4) *Stop if* $||\mathbf{u}^{(k)} - \mathbf{u}^{(k-1)}||_{L^2_\nu(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} < tol$ *and return* $\mathbf{u}^{(k)}$ *as the approximate eigenvector.*

Once the approximate eigenvector $\mathbf{u}^{(k)} \in W_{\mathcal{A}}^N$ has been computed, the corresponding eigenvalue $\mu^{(k)} \in W_{\mathcal{A}}$ may be evaluated from the Rayleigh quotient, as in [11], or alternatively from the linear system

$$P_{\mathcal{A}}(s\mu^{(k)}) = 1. \tag{37}$$

Lemma 1 guarantees the invertibility of the linear system (34) and, assuming that $s(y) > 0$ for all $y \in \Gamma$, the invertibility of the systems (36) and (37). The nonlinear system (35) may be solved using for instance Newton's method.

Remark 2 For the computation of non-extremal eigenmodes, one may proceed as in [11] and replace $\mathbf{K}(y)$ in (34) with $(\mathbf{K}(y) - \lambda\mathbf{M})$, where $\lambda \in \mathbb{R}$ is a suitably chosen parameter. In this case we expect the algorithm to converge to an eigenpair for which the eigenvalue is close to λ. Note, however, that now the existence of a unique solution to (34) is not necessarily guaranteed by Lemma 1.

We try to write Algorithm 1 in a computationally more convenient form. The projections in the algorithm can be computed explicitly using the notation introduced in

Sect. 4. It is easy to verify that Eqs. (34)–(36) become

$$\sum_{m=0}^{\infty}\sum_{\beta\in\mathcal{A}}\mathbf{K}^{(m)}\mathbf{v}_\beta c_{m\alpha\beta}=\mathbf{M}\mathbf{u}_\alpha^{(k-1)}\quad\forall\alpha\in\mathcal{A},\tag{38}$$

$$\sum_{\beta\in\mathcal{A}}\sum_{\gamma\in\mathcal{A}}s_\beta s_\gamma c_{\alpha\beta\gamma}=\sum_{\beta\in\mathcal{A}}\sum_{\gamma\in\mathcal{A}}\langle\mathbf{v}_\beta,\mathbf{v}_\gamma\rangle_{\mathbb{R}^N_{\mathbf{M}}}c_{\alpha\beta\gamma}\quad\forall\alpha\in\mathcal{A},\tag{39}$$

$$\sum_{\beta\in\mathcal{A}}\sum_{\gamma\in\mathcal{A}}s_\beta\mathbf{u}_\gamma^{(k)}c_{\alpha\beta\gamma}=\mathbf{v}_\alpha\quad\forall\alpha\in\mathcal{A}\tag{40}$$

respectively. Given $\hat{s}=\{s_\alpha\}_{\alpha\in\mathcal{A}}\in\mathbb{R}^P$ we define matrices

$$\begin{aligned}\Delta(\hat{s})&:=\sum_{\alpha\in\mathcal{A}}G^{(\alpha)}s_\alpha,\\ \widehat{\mathbf{K}}&:=\sum_{m=0}^{M(\mathcal{A})}G^{(m)}\otimes\mathbf{K}^{(m)},\\ \widehat{\mathbf{M}}&:=I_P\otimes\mathbf{M},\\ \mathbf{S}&:=\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{K}},\\ \mathbf{T}(\hat{s})&:=\Delta(\hat{s})\otimes\mathbf{I}_N,\end{aligned}$$

where $M(\mathcal{A}):=\max\{m\in\mathbb{N}\mid\exists\alpha\in\mathcal{A}\text{ s.t. }\alpha_m\neq 0\}$ and $I_P\in\mathbb{R}^{P\times P}$ and $\mathbf{I}_N\in\mathbb{R}^{N\times N}$ are identity matrices. We also define the nonlinear function $F:\mathbb{R}^P\times\mathbb{R}^{PN}\to\mathbb{R}^P$ via

$$F_\alpha(\hat{s},\hat{\mathbf{v}}):=\hat{s}\cdot G^{(\alpha)}\hat{s}-\hat{\mathbf{v}}\cdot(G^{(\alpha)}\otimes\mathbf{M})\hat{\mathbf{v}},\quad\alpha\in\mathcal{A}$$

and let $F^s:\mathbb{R}^P\times\mathbb{R}^P\to\mathbb{R}^P$ and $F^v:\mathbb{R}^{PN}\times\mathbb{R}^{PN}\to\mathbb{R}^P$ denote the associated bilinear forms given by $F^s_\alpha(\hat{s},\hat{t}):=\hat{s}\cdot G^{(\alpha)}\hat{t}$ and $F^v_\alpha(\hat{\mathbf{v}},\hat{\mathbf{w}}):=\hat{\mathbf{v}}\cdot(G^{(\alpha)}\otimes\mathbf{M})\hat{\mathbf{w}}$. Now Algorithm 1 may be rewritten in the following form.

Algorithm 2 (Spectral inverse iteration in tensor form) *Fix* $tol>0$ *and let* $\hat{\mathbf{u}}^{(0)}=\{u^{(0)}_{\alpha i}\}_{\alpha\in\mathcal{A},i\in J}\in\mathbb{R}^{PN}$ *be an initial guess for the eigenvector. For* $k=1,2,\ldots$ *do*

(1) *Solve* $\hat{\mathbf{v}}=\{v_{\alpha i}\}_{\alpha\in\mathcal{A},i\in J}\in\mathbb{R}^{PN}$ *from the linear system*

$$\widehat{\mathbf{K}}\hat{\mathbf{v}}=\widehat{\mathbf{M}}\hat{\mathbf{u}}^{(k-1)}.\tag{41}$$

(2) *Solve* $\hat{s}=\{s_\alpha\}_{\alpha\in\mathcal{A}}\in\mathbb{R}^P$ *from the nonlinear system*

$$F(\hat{s},\hat{\mathbf{v}})=0\tag{42}$$

with the initial guess $s_\alpha=||\hat{\mathbf{v}}||_{\mathbb{R}^P\otimes\mathbb{R}^N_{\mathbf{M}}}\delta_{\alpha 0}$ *for* $\alpha\in\mathcal{A}$.

(3) *Solve* $\hat{\mathbf{u}}^{(k)} = \{u_{\alpha i}^{(k)}\}_{\alpha\in\mathcal{A}, i\in J} \in \mathbb{R}^{PN}$ *from the linear system*

$$\mathbf{T}(\hat{s})\hat{\mathbf{u}}^{(k)} = \hat{\mathbf{v}}. \tag{43}$$

(4) *Stop if* $||\hat{\mathbf{u}}^{(k)} - \hat{\mathbf{u}}^{(k-1)}||_{\mathbb{R}^P\otimes\mathbb{R}^N_{\mathbf{M}}} < tol$ *and return* $\hat{\mathbf{u}}^{(k)}$ *as the approximate eigenvector.*

The approximate eigenvalue $\hat{\mu}^{(k)} \in \mathbb{R}^P$ may now be solved from the equation

$$\Delta(\hat{s})\hat{\mu}^{(k)} = \hat{e}_1, \tag{44}$$

where $\hat{e}_1 = \{\delta_{\alpha 0}\}_{\alpha\in\mathcal{A}} \in \mathbb{R}^P$.

Remark 3 In [11] Newton's method with the initial guess $s_\alpha = ||\mathbf{v}_\alpha||_{\mathbb{R}^N_{\mathbf{M}}}$ was suggested for the system of Eq. (42). Here the initial guess is somewhat different and corresponds to $s_0 = ||\mathbf{v}||_{L^2_\nu(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}}$ (and $s_\alpha = 0$ for $\alpha \neq 0$).

In general it is not guaranteed that the Newton iteration for the system (42) converges to a solution. The following proposition will give some insight to the conditions under which this happens to be the case.

Proposition 3 *Fix* $\hat{\mathbf{v}} \in \mathbb{R}^{PN}$ *and let* $\hat{s}^{(0)} = \{s_\alpha^{(0)}\}_{\alpha\in\mathcal{A}} \in \mathbb{R}^P$ *be given by* $s_\alpha^{(0)} = ||\hat{\mathbf{v}}||_{\mathbb{R}^P\otimes\mathbb{R}^N_{\mathbf{M}}}\delta_{\alpha 0}$ *for* $\alpha \in \mathcal{A}$. *Assume that there is a norm* $||\cdot||_*$ *on* $\mathbb{R}^P$ *and* $r > 0$ *such that*

$$||F^s(\hat{s}, \hat{t})||_* \leq C_F||\hat{s}||_*||\hat{t}||_*$$

for all $\hat{s}, \hat{t}$ *in* $B(\hat{s}^{(0)}, r) := \{\hat{s} \in \mathbb{R}^P \mid ||\hat{s} - \hat{s}^{(0)}||_* \leq r\}$. *If*

$$||F(\hat{s}^{(0)}, \hat{\mathbf{v}})||_* < C_F^{-1}||\hat{\mathbf{v}}||^2_{\mathbb{R}^P\otimes\mathbb{R}^N_{\mathbf{M}}}$$

then the Newton method for $F(\cdot, \hat{\mathbf{v}}) = 0$ *with the initial guess* $\hat{s}^{(0)}$ *converges to a unique solution in* $B(\hat{s}^{(0)}, r)$.

Proof This is a direct application of the Newton-Kantorovich theorem for the equation $F(\cdot, \hat{\mathbf{v}}) = 0$, see [13] (Theorem 6, 1.XVIII). Note that the first derivative (Jacobian) of $F(\cdot, \hat{\mathbf{v}})$ at $\hat{s}^{(0)}$ is $2||\hat{\mathbf{v}}||_{\mathbb{R}^P\otimes\mathbb{R}^N_{\mathbf{M}}} I_P$ and the second derivative is represented by the tensor of coefficients $2c_{\alpha\beta\gamma}$. □

From Proposition 3 we see that convergence of the Newton iteration is a consequence of the boundedness of the function F^s, which again is ultimately determined by the structure of the multi-index set $\mathcal{A}$.

5.2 Analysis of convergence

Due to a lack of general mathematical theory for multi-parametric eigenvalue problems we rely on a slightly unconventional approach in analyzing our algorithm. First of all, we restrict ourselves to asymptotic analysis since the underlying problem is nonlinear and thus hard to analyze globally. Second, we will analyze the solutions pointwise in the parameter space and deduce convergence theorems from classical eigenvalue perturbation bounds.

5.2.1 Characterization of the dominant fixed point

The classical inverse iteration converges to the dominant eigenpair of the inverse matrix. In a somewhat similar fashion the spectral inverse iteration tends to converge to a certain fixed point, which we shall refer to as the dominant fixed point. Here we will establish a connection between this dominant fixed point of the spectral inverse iteration and the dominant eigenpair of the inverse of the parametric matrix under consideration. This connection is obtained by considering the fixed point as a pointwise perturbation of the eigenvalue problem of the parametric matrix.

If $\mathbf{u}_{\mathcal{A}} \in W_{\mathcal{A}}^N$ is a fixed point of the Algorithm 1, then there exists a pair $(s, \mathbf{v}) \in W_{\mathcal{A}} \times W_{\mathcal{A}}^N$ such that $\mathbf{u}_{\mathcal{A}} = \mathbf{M}^{-1} P_{\mathcal{A}}(\mathbf{K}\mathbf{v})$ and

$$\begin{cases} P_{\mathcal{A}}\left(s P_{\mathcal{A}}(\mathbf{K}\mathbf{v})\right) = \mathbf{M}\mathbf{v} \\ P_{\mathcal{A}}\left(s^2 - ||\mathbf{v}||^2_{\mathbb{R}^N_{\mathbf{M}}}\right) = 0. \end{cases} \tag{45}$$

We call $\mathbf{u}_{\mathcal{A}}$ the dominant fixed point if, whenever $(\tilde{s}, \tilde{\mathbf{v}}) \neq (s, \mathbf{v})$ also solves the system (45), then $s(y) > \tilde{s}(y)$ for all $y \in \Gamma$. For any fixed $y \in \Gamma$ we may write (45) as

$$\begin{cases} s(y)\mathbf{K}(y)\mathbf{v}(y) = \mathbf{M}\mathbf{v}(y) + s(y)R_{\mathcal{A}}(\mathbf{K}\mathbf{v})(y) + R_{\mathcal{A}}(s P_{\mathcal{A}}(\mathbf{K}\mathbf{v}))(y) \\ s^2(y) = ||\mathbf{v}(y)||^2_{\mathbb{R}^N_{\mathbf{M}}} + R_{\mathcal{A}}\left(s^2 - ||\mathbf{v}||^2_{\mathbb{R}^N_{\mathbf{M}}}\right)(y). \end{cases} \tag{46}$$

The following Lemma will be helpful in establishing a connection between the eigenpair of interest and the system (46).

Lemma 2 *Denote by $||\cdot||$ the standard Euclidean norm on $\mathbb{R}^N$. Assume that $S \in \mathbb{R}^{N\times N}$ can be diagonalized as*

$$S(x\ X) = (x\ X)\begin{pmatrix} \lambda_1 & 0 \\ 0 & \Lambda \end{pmatrix}, \tag{47}$$

where $\lambda_1 \in \mathbb{R}$, $\Lambda = \mathrm{diag}(\lambda_2, \ldots, \lambda_N)$ is real, and $(x\ X)$ is orthogonal. Assume also that $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_N$ and denote $\hat{\lambda} := \lambda_1 - \lambda_2$. Let $\rho \in \mathbb{R}$ and $r \in \mathbb{R}^N$ be such that $|\rho| \leq 1/2$ and $||r|| \leq \hat{\lambda}/8$. Then there exist $\kappa \geq 1/2$ and $\pi \in \mathbb{R}^{N-1}$ such that

(i) *The pair (s, w) given by $s = \lambda_1 - \kappa^{-1}x^T r$ and $w = \kappa x + X\pi$ solves the system*

$$\begin{cases} Sw = sw + r \\ ||w||^2 = 1 + \rho. \end{cases} \tag{48}$$

(ii) *If* $(\tilde{s}, \tilde{w}) \neq (s, w)$ *also solves the system* (48)*, then* $s > \tilde{s}$ *or* $x^T \tilde{w} < 0$.
(iii) *There exists* $C > 0$ *such that* $|\kappa - 1| \leq C(|\rho| + \hat{\lambda}^{-2}||r||^2)$ *and* $||\pi|| \leq C\hat{\lambda}^{-1}||r||$.

Proof (i) Let $s(\kappa) = \lambda_1 - \kappa^{-1} x^T r$. For any $\kappa \geq 1/2$ we have $|\kappa^{-1} x^T r| \leq \hat{\lambda}/4$ so that

$$\min_{2 \leq i \leq N} |\lambda_i - s(\kappa)| = \min_{2 \leq i \leq N} |\lambda_1 - \lambda_i - \kappa^{-1} x^T r| \geq \hat{\lambda} - \frac{1}{4}\hat{\lambda} > \frac{1}{2}\hat{\lambda} \tag{49}$$

and

$$||(\Lambda - s(\kappa) I)^{-1}|| \leq 2\hat{\lambda}^{-1}. \tag{50}$$

The function

$$f(\kappa) = \kappa^2 + ||(\Lambda - s(\kappa) I)^{-1} X^T r||^2 - 1 - \rho \tag{51}$$

is strictly increasing for $\kappa \geq 1/2$ since

$$\begin{aligned} \kappa^2 f'(\kappa) &= 2\kappa^3 + 2x^T r ||(\Lambda - s(\kappa) I)^{-\frac{3}{2}} X^T r||^2 \\ &\geq 2(\kappa^3 - (2\hat{\lambda}^{-1})^3 ||r||^3) \\ &> 2(\kappa^3 - 2^{-3}) \geq 0. \end{aligned} \tag{52}$$

One may also verify that $f(1/2) < 0$ and $f(2) > 0$. Thus, we may choose $\kappa > 1/2$ such that $f(\kappa) = 0$. For $w = \kappa x + X\pi$ we obtain

$$Sw - sw = \kappa Sx + SX\pi - \kappa s x - sX\pi = \kappa(\lambda_1 - s)x + X(\Lambda - sI)\pi \tag{53}$$

so the equation $Sw = sw + r$ is equivalent to

$$\begin{cases} x^T(Sw - sw - r) = \kappa(\lambda_1 - s) - x^T r = 0 \\ X^T(Sw - sw - r) = (\Lambda - sI)\pi - X^T r = 0. \end{cases} \tag{54}$$

Choosing $s = s(\kappa)$ and $\pi = (\Lambda - sI)^{-1} X^T r$ we see that both equations are satisfied. Moreover

$$||w||^2 = \kappa^2 + ||\pi||^2 = f(\kappa) + 1 + \rho = 1 + \rho. \tag{55}$$

(ii) Suppose $(\tilde{s}, \tilde{w})$ also solves the system (48) and write $\tilde{w} = \tilde{\kappa} x + X\tilde{\pi}$ for some $\tilde{\kappa} \in \mathbb{R}$ and $\tilde{\pi} \in \mathbb{R}^{N-1}$. In the nontrivial case we have $\tilde{\kappa} = x^T \tilde{w} > 0$. Assume first that $0 \leq \tilde{\kappa} \leq 1/2$. We have

$$\begin{aligned} \tilde{s} &= \frac{\tilde{w}^T S \tilde{w} - \tilde{w}^T r}{||\tilde{w}||^2} = \frac{\lambda_1 \tilde{\kappa}^2 + \tilde{\pi}^T \Lambda \tilde{\pi} - \tilde{w}^T r}{||\tilde{w}||^2} \leq \frac{\lambda_1 \tilde{\kappa}^2 + \lambda_2 ||\tilde{\pi}||^2 + ||\tilde{w}||||r||}{||\tilde{w}||^2} \\ &= \lambda_2 + \frac{\tilde{\kappa}^2}{1 + \rho}\hat{\lambda} + \frac{||r||}{(1 + \rho)^{\frac{1}{2}}}. \end{aligned} \tag{56}$$

Since $s \geq \lambda_1 - \kappa^{-1}||r||$, we deduce that

$$s - \tilde{s} \geq \hat{\lambda} - \kappa^{-1}||r|| - \frac{\tilde{\kappa}^2}{1+\rho}\hat{\lambda} - \frac{||r||}{(1+\rho)^{\frac{1}{2}}} > \left(1 - \frac{1}{4} - \frac{1}{2} - \frac{\sqrt{2}}{8}\right)\hat{\lambda} > 0. \quad (57)$$

Now let $\tilde{\kappa} \geq 1/2$. If $(\tilde{s}, \tilde{w})$ is to solve (48) then, as in part (i), we should have

$$\begin{cases} \tilde{\kappa}(\lambda_1 - \tilde{s}) - x^T r = 0 \\ (\Lambda - \tilde{s}I)\tilde{\pi} - X^T r = 0. \end{cases} \quad (58)$$

From the first equation we obtain $\tilde{s} = \lambda_1 - \tilde{\kappa}^{-1}x^T r$. Due to $|\tilde{\kappa}^{-1}x^T r| \leq \hat{\lambda}/4$ the matrix $(\Lambda - \tilde{s}I)$ is invertible so the second equation gives $\tilde{\pi} = (\Lambda - \tilde{s}I)^{-1}X^T r$. Here $\tilde{\kappa} \geq 1/2$ must be chosen so that $f(\tilde{\kappa}) = 0$ and therefore $(\tilde{s}, \tilde{w}) = (s, w)$.

(iii) From $f(\kappa) = 0$ and $\kappa \geq 1/2$ we deduce that

$$|\kappa - 1| \leq (\kappa+1)^{-1}(|\rho| + ||(\Lambda - s(\kappa)I)^{-1}X^T r||^2) \leq |\rho| + 4\hat{\lambda}^{-2}||r||^2 \quad (59)$$

and

$$||\pi|| = ||(\Lambda - s(\kappa)I)^{-1}X^T r|| \leq 2\hat{\lambda}^{-1}||r||. \quad (60)$$

Thus, the claim follows. □

Applying Lemma 2 to the system (46) pointwise for $y \in \Gamma$ we obtain the following result.

Proposition 4 *Let $\mathbf{u}_{\mathcal{A}} \in W_{\mathcal{A}}^N$ be the dominant fixed point of Algorithm 1 and denote by $(s, \mathbf{v})$ the associated pair in $W_{\mathcal{A}} \times W_{\mathcal{A}}^N$ that solves (45). Let $\mu_{\mathcal{A}} \in W_{\mathcal{A}}$ be such that $P_{\mathcal{A}}(s\mu_{\mathcal{A}}) = 1$. For $y \in \Gamma$ denote by $\hat{\lambda}(y)$ the gap between the two largest eigenvalues of $\mathbf{K}^{-1}(y)\mathbf{M}$. Assume that $\inf_{y\in\Gamma} s(y) > 0$ and $\inf_{y\in\Gamma} \hat{\lambda}(y) > 0$. For $y \in \Gamma$ define*

$$\mathbf{r}(y) := \mathbf{K}^{-1}(y)R_{\mathcal{A}}(\mathbf{K}\mathbf{v})(y) + s^{-1}(y)\mathbf{K}^{-1}(y)R_{\mathcal{A}}(sP_{\mathcal{A}}(\mathbf{K}\mathbf{v}))(y)$$

and

$$\rho(y) := s^{-2}(y)R_{\mathcal{A}}\left(s^2 - ||\mathbf{v}||^2_{\mathbb{R}^N_{\mathbf{M}}}\right)(y).$$

If

$$r_* := \sup_{y\in\Gamma} \hat{\lambda}^{-1}(y)||\mathbf{r}(y)||_{\mathbb{R}^N_{\mathbf{M}}} < \frac{1}{8}$$

and

$$\rho_* := \sup_{y\in\Gamma} |\rho(y)| < \frac{1}{2}$$

then there exists $C > 0$ such that

$$|\mu_{\mathcal{A}}(y) - \mu_h(y)| \leq C \left(\max\{\mu_h^2(y), s^{-2}(y)\} ||\mathbf{r}(y)||_{\mathbb{R}^N_{\mathbf{M}}} + s^{-1}(y) |R_{\mathcal{A}}(s\mu_{\mathcal{A}})(y)| \right) \tag{61}$$

and

$$\begin{aligned} &||\mathbf{u}_{\mathcal{A}}(y) - \mathbf{u}_h(y)||_{\mathbb{R}^N_{\mathbf{M}}} \\ &\quad \leq C \left(|\rho(y)| + \hat{\lambda}^{-1}(y) ||\mathbf{r}(y)||_{\mathbb{R}^N_{\mathbf{M}}} + s^{-1}(y) ||\mathbf{M}^{-1} R_{\mathcal{A}}(s P_{\mathcal{A}}(\mathbf{K}\mathbf{v}))(y)||_{\mathbb{R}^N_{\mathbf{M}}} \right), \end{aligned} \tag{62}$$

where $\mu_h : \Gamma \to \mathbb{R}$ is the smallest eigenvalue of $\mathbf{M}^{-1}\mathbf{K}(y)$ and $\mathbf{u}_h : \Gamma \to \mathbb{R}^N$ is the corresponding eigenvector normalized in $||\cdot||_{\mathbb{R}^N_{\mathbf{M}}}$ (and with appropriate sign).

Proof It is easy to see that the system (46) is equivalent to

$$\begin{cases} \mathbf{M}^{\frac{1}{2}} \mathbf{K}^{-1}(y) \mathbf{M}^{\frac{1}{2}} \mathbf{w}(y) = s(y)\mathbf{w}(y) - \mathbf{M}^{\frac{1}{2}} \mathbf{r}(y) \\ ||\mathbf{w}(y)||^2_{\mathbb{R}^N} = 1 - \rho(y), \end{cases} \tag{63}$$

where $\mathbf{w}(y) = s^{-1}(y)\mathbf{M}^{\frac{1}{2}}\mathbf{v}(y)$. By Lemma 2 the solution with the pointwise largest $s(y)$ can be written as

$$\begin{cases} s(y) = \lambda_1(y) + \kappa^{-1}(y) \mathbf{x}^T(y) \mathbf{M}^{\frac{1}{2}} \mathbf{r}(y) \\ \mathbf{w}(y) = \kappa(y)\mathbf{x}(y) + \mathbf{X}(y)\boldsymbol{\pi}(y), \end{cases} \tag{64}$$

where $\kappa : \Gamma \to [1/2, \infty)$ and $\boldsymbol{\pi} : \Gamma \to \mathbb{R}^{N-1}$ are such that

$$|\kappa(y) - 1|^2 + ||\boldsymbol{\pi}(y)||^2_{\mathbb{R}^N} \leq C \left(|\rho(y)| + \hat{\lambda}^{-2}(y) ||\mathbf{r}(y)||^2_{\mathbb{R}^N_{\mathbf{M}}} \right)^2 + C\hat{\lambda}^{-2}(y) ||\mathbf{r}(y)||^2_{\mathbb{R}^N_{\mathbf{M}}},$$

$\lambda_1(y) = \mu_h^{-1}(y)$ is the pointwise largest eigenvalue of $\mathbf{S}(y)$ and $\mathbf{x}(y) = \mathbf{M}^{\frac{1}{2}}\mathbf{u}_h(y)$ is the corresponding eigenvector. The matrix $(\mathbf{x}(y)\ \mathbf{X}(y))$ is orthonormal for every $y \in \Gamma$. A Taylor expansion of $s^{-1}(y)$ yields

$$\begin{aligned} |s^{-1}(y) - \mu_h(y)| &= C(\mu_h^{-1}(y) + \xi(y))^{-2} ||\mathbf{r}(y)||_{\mathbb{R}^N_{\mathbf{M}}} \\ &\leq C \max\{\mu_h^2(y), s^{-2}(y)\} ||\mathbf{r}(y)||_{\mathbb{R}^N_{\mathbf{M}}}, \end{aligned} \tag{65}$$

where $\xi(y)$ is such that $0 \leq \xi(y) \leq \kappa^{-1}(y)\mathbf{x}^T(y)\mathbf{M}^{\frac{1}{2}}\mathbf{r}(y)$. Combining this with the equation

$$\mu_{\mathcal{A}}(y) = s^{-1}(y) + s^{-1}(y) R_{\mathcal{A}}(s\mu_{\mathcal{A}})(y) \tag{66}$$

obtained from the condition $P_{\mathcal{A}}(s\mu_{\mathcal{A}}) = 1$, we have altogether that

$$|\mu_{\mathcal{A}}(y) - \mu_h(y)| \leq C \max\{\mu_h^2(y), s^{-2}(y)\} ||\mathbf{r}(y)||_{\mathbb{R}^N_{\mathbf{M}}} + s^{-1}(y) |R_{\mathcal{A}}(s\mu_{\mathcal{A}})(y)|. \tag{67}$$

Furthermore,

$$\mathbf{M}^{\frac{1}{2}}\mathbf{u}_{\mathcal{A}}(y) = \mathbf{M}^{-\frac{1}{2}}P_{\mathcal{A}}(\mathbf{K}\mathbf{v})(y) = s^{-1}(y)\mathbf{M}^{-\frac{1}{2}}R_{\mathcal{A}}(sP_{\mathcal{A}}(\mathbf{K}\mathbf{v}))(y) + \mathbf{w}(y) \tag{68}$$

from which it follows that

$$\begin{aligned}
||\mathbf{u}_{\mathcal{A}}(y) - \mathbf{u}_h(y)||^2_{\mathbb{R}^N_{\mathbf{M}}} &= ||\mathbf{M}^{\frac{1}{2}}\mathbf{u}_{\mathcal{A}}(y) - \mathbf{w}(y)||^2_{\mathbb{R}^N} + ||\mathbf{w}(y) - \mathbf{M}^{\frac{1}{2}}\mathbf{u}_h(y)||^2_{\mathbb{R}^N} \\
&= ||s^{-1}(y)\mathbf{M}^{-\frac{1}{2}}R_{\mathcal{A}}(sP_{\mathcal{A}}(\mathbf{K}\mathbf{v}))(y)||^2_{\mathbb{R}^N} + ||(\kappa(y)-1)\mathbf{x}(y) + \mathbf{X}(y)\boldsymbol{\pi}(y)||^2_{\mathbb{R}^N} \\
&= s^{-1}(y)||\mathbf{M}^{-1}R_{\mathcal{A}}(sP_{\mathcal{A}}(\mathbf{K}\mathbf{v}))(y)||^2_{\mathbb{R}^N_{\mathbf{M}}} + |\kappa(y)-1|^2 + ||\boldsymbol{\pi}(y)||^2_{\mathbb{R}^N} \\
&\leq C\left(|\rho(y)| + \hat{\lambda}^{-1}(y)||\mathbf{r}(y)||_{\mathbb{R}^N_{\mathbf{M}}} + s^{-1}(y)||\mathbf{M}^{-1}R_{\mathcal{A}}(sP_{\mathcal{A}}(\mathbf{K}\mathbf{v}))(y)||_{\mathbb{R}^N_{\mathbf{M}}}\right)^2.
\end{aligned} \tag{69}$$

This concludes the proof. □

Remark 4 Note that we have not proven the existence of a dominant fixed point of the Algorithm 1. The residuals $\mathbf{r}$ and ρ in Proposition 4 depend on the pair $(s, \mathbf{v}) \in W_{\mathcal{A}} \times W^N_{\mathcal{A}}$ and hence Lemma 2 by itself is not sufficient to guarantee the existence of a dominant fixed point.

5.2.2 Convergence of the dominant fixed point to a parametric eigenpair

The next step in our analysis is to bound the error between the dominant fixed point of Algorithm 1 and the dominant eigenpair of the inverse of the parametric matrix. To this end we will use the pointwise estimate obtained previously.

From Proposition 4 we may easily deduce the following result.

Theorem 2 *Let $\mathbf{u}_{\mathcal{A}} \in W^N_{\mathcal{A}}$ be the dominant fixed point of Algorithm 1 and denote by $(s, \mathbf{v})$ the associated pair in $W_{\mathcal{A}} \times W^N_{\mathcal{A}}$ that solves (45). Let $\mu_{\mathcal{A}} \in W_{\mathcal{A}}$ be such that $P_{\mathcal{A}}(s\mu_{\mathcal{A}}) = 1$. For $y \in \Gamma$ denote by $\hat{\lambda}(y)$ the gap between the two largest eigenvalues of $\mathbf{K}^{-1}(y)\mathbf{M}$. Assume that $s_* := \inf_{y\in\Gamma} s(y) > 0$, $\hat{\lambda}_* := \inf_{y\in\Gamma}\hat{\lambda}(y) > 0$ and that the quantity*

$$||R_{\mathcal{A}}(\mathbf{K}\mathbf{v})||_{L^\infty(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} + ||R_{\mathcal{A}}\left(sP_{\mathcal{A}}(\mathbf{K}\mathbf{v})\right)||_{L^\infty(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} + \left|\left|R_{\mathcal{A}}\left(s^2 - ||\mathbf{v}||^2_{\mathbb{R}^N_{\mathbf{M}}}\right)\right|\right|_{L^\infty(\Gamma)}$$

is small enough. Then there exists $C > 0$ such that

$$\begin{aligned}
||\mu_{\mathcal{A}} - \mu_h||_{L^2_\nu(\Gamma)} \leq C\Big(&||R_{\mathcal{A}}(\mathbf{K}\mathbf{v})||_{L^2_\nu(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} \\
&+ ||R_{\mathcal{A}}(sP_{\mathcal{A}}(\mathbf{K}\mathbf{v}))||_{L^2_\nu(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} + ||R_{\mathcal{A}}\left(s\mu_{\mathcal{A}}\right)||_{L^2_\nu(\Gamma)}\Big)
\end{aligned} \tag{70}$$

and

$$||\mathbf{u}_{\mathcal{A}} - \mathbf{u}_h||_{L^2_v(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} \leq C \left(||R_{\mathcal{A}}(\mathbf{Kv})||_{L^2_v(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} + ||R_{\mathcal{A}}(s\,P_{\mathcal{A}}(\mathbf{Kv}))||_{L^2_v(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} + \left|\left| R_{\mathcal{A}}\left(s^2 - ||\mathbf{v}||^2_{\mathbb{R}^N_{\mathbf{M}}}\right)\right|\right|_{L^2_v(\Gamma)} \right), \tag{71}$$

where $\mu_h : \Gamma \to \mathbb{R}$ *is the smallest eigenvalue of* $\mathbf{M}^{-1}\mathbf{K}(y)$ *and* $\mathbf{u}_h : \Gamma \to \mathbb{R}^N$ *is the corresponding eigenvector normalized in* $||\cdot||_{\mathbb{R}^N_{\mathbf{M}}}$ *(and with appropriate sign). Here* C *depends only on* s_*, $\hat{\lambda}_*$, $\mu_h^* := \sup_{y\in\Gamma} \mu_h(y)$, $K_* = \sup_{y\in\Gamma} ||\mathbf{K}^{-1}(y)||_{\mathbb{R}^N_{\mathbf{M}}}$, *and* $M_* = ||\mathbf{M}^{-1}||_{\mathbb{R}^N_{\mathbf{M}}}$.

Proof With $\mathbf{r}$ defined as in Proposition 4 we have

$$||\mathbf{r}||_{L^2_v(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} \leq K_* \left(||R_{\mathcal{A}}(\mathbf{Kv})||_{L^2_v(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} + s_*^{-1} ||R_{\mathcal{A}}(s\,P_{\mathcal{A}}(\mathbf{Kv}))||_{L^2_v(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} \right), \tag{72}$$

and

$$||\mathbf{M}^{-1} R_{\mathcal{A}}(s\,P_{\mathcal{A}}(\mathbf{Kv}))(y)||_{L^2_v(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} \leq M_* ||R_{\mathcal{A}}(s\,P_{\mathcal{A}}(\mathbf{Kv}))(y)||_{L^2_v(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}}. \tag{73}$$

The bounds (70) and (71) now follow from Proposition 4. □

By Proposition 1 the exact eigenvalue and eigenvector of problem (9) are analytic functions of the parameter vector $y \in \Gamma$. This suggests that the residuals on the right hand side of Eqs. (70) and (71) can be asymptotically estimated from Proposition 2.

5.2.3 Convergence of the spectral inverse iteration to the dominant fixed point

The classical inverse iteration converges to the dominant eigenpair of the inverse matrix at a speed characterized by the gap between the two largest eigenvalues. Here we will establish a similar asymptotic result for the convergence of the spectral inverse iteration towards the dominant fixed point.

Fixed points of the spectral inverse iteration may be characterized using the tensor notation of Algorithm 2. Let $\hat{\mathbf{u}}_{\mathcal{A}} \in \mathbb{R}^{PN}$ be a fixed point of the algorithm, i.e., $\hat{\mathbf{u}}_{\mathcal{A}} = \mathbf{S}\hat{\mathbf{v}}$ and $(\hat{s}, \hat{\mathbf{v}}) \in \mathbb{R}^P \times \mathbb{R}^{PN}$ are such that

$$\begin{cases} \hat{\mathbf{v}} = \mathbf{T}(\hat{s})\mathbf{S}\hat{\mathbf{v}} \\ F(\hat{s}, \hat{\mathbf{v}}) = 0. \end{cases} \tag{74}$$

Define a linear operator $\mathbf{R}(\hat{s}, \hat{\mathbf{v}}) : \mathbb{R}^{PN} \to \mathbb{R}^{PN}$ by

$$\mathbf{R}(\hat{s}, \hat{\mathbf{v}})\hat{\mathbf{w}} := \hat{\mathbf{w}} - \mathbf{T}\left(\Delta^{-1}(\hat{s}) F^v(\hat{\mathbf{v}}, \hat{\mathbf{w}})\right) \mathbf{T}^{-1}(\hat{s})\hat{\mathbf{v}}.$$

The convergence of the spectral inverse iteration to the fixed point $\hat{\mathbf{u}}_{\mathcal{A}}$ can now be related to the ratio of the norms of $\Delta^{-1}(\hat{s})$ and $\mathbf{R}(\hat{s}, \hat{\mathbf{v}})\mathbf{S}^{-1}$.

Theorem 3 *Let $\hat{\mathbf{u}}_{\mathcal{A}} \in \mathbb{R}^{PN}$ be a fixed point of the Algorithm 2 and $(\hat{s}, \hat{\mathbf{v}}) \in \mathbb{R}^P \times \mathbb{R}^{PN}$ a corresponding solution to (74). Assume that $\Delta(\hat{s})$ is invertible. Let $\hat{\mu}_{\mathcal{A}} = \Delta^{-1}(\hat{s})\hat{e}_1$, where $\hat{e}_1 = \{\delta_{\alpha 0}\}_{\alpha \in \mathcal{A}} \in \mathbb{R}^P$. Set $\phi_{\min} := ||\Delta^{-1}(\hat{s})||^{-1}_{\mathbb{R}^P}$ and $\psi_{\max} := ||\mathbf{R}(\hat{s}, \hat{\mathbf{v}})\mathbf{S}^{-1}||_{\mathbb{R}^P \otimes \mathbb{R}^N_{\mathbf{M}}}$. Then for any $\varepsilon > 0$ the iterates of Algorithm 2 satisfy*

$$||\hat{\mathbf{u}}^{(k)} - \hat{\mathbf{u}}_{\mathcal{A}}||_{\mathbb{R}^P \otimes \mathbb{R}^N_{\mathbf{M}}} \leq \left(\frac{\psi_{\max}}{\phi_{\min}} + \varepsilon \right) ||\hat{\mathbf{u}}^{(k-1)} - \hat{\mathbf{u}}_{\mathcal{A}}||_{\mathbb{R}^P \otimes \mathbb{R}^N_{\mathbf{M}}}, \quad k \in \mathbb{N} \tag{75}$$

whenever $\hat{\mathbf{u}}^{(k)}$ is sufficiently close to $\hat{\mathbf{u}}_{\mathcal{A}}$. Furthermore, there exists $C > 0$ such that

$$||\hat{\mu}^{(k)} - \hat{\mu}_{\mathcal{A}}||_{\mathbb{R}^P} \leq C||\hat{\mathbf{u}}^{(k)} - \hat{\mathbf{u}}_{\mathcal{A}}||_{\mathbb{R}^P \otimes \mathbb{R}^N_{\mathbf{M}}}, \quad k \in \mathbb{N}. \tag{76}$$

Proof The partial derivative (Jacobian) of the function $F(\hat{s} + \hat{t}, \hat{\mathbf{v}} + \hat{\mathbf{w}})$ with respect to $\hat{t}$ at $\hat{t} = 0$ is given by $2\Delta(\hat{s})$. The implicit function theorem now guarantees that there is a unique differentiable function $\hat{t}(\hat{\mathbf{w}})$ defined in a neighbourhood of $\hat{\mathbf{w}} = 0$ such that $F(\hat{s} + \hat{t}(\hat{\mathbf{w}}), \hat{\mathbf{v}} + \hat{\mathbf{w}}) = 0$. Computing the first order approximation of this function we see that for $\hat{\mathbf{w}}$ small enough

$$\hat{t}(\hat{\mathbf{w}}) = \Delta^{-1}(\hat{s})F^v(\hat{\mathbf{v}}, \hat{\mathbf{w}}) + \text{h.o.t. in } \hat{\mathbf{w}}, \tag{77}$$

where h.o.t. stands for higher order terms. From (77) we obtain

$$\begin{aligned} \mathbf{T}^{-1}\left(\hat{s} + \hat{t}(\mathbf{w})\right) &= \left(\mathbf{T}(\hat{s}) + \mathbf{T}\left(\Delta^{-1}(\hat{s})F^v(\hat{\mathbf{v}}, \hat{\mathbf{w}})\right) + \text{h.o.t. in } \hat{\mathbf{w}}\right)^{-1} \\ &= \mathbf{T}^{-1}(\hat{s}) - \mathbf{T}^{-1}(\hat{s})\mathbf{T}\left(\Delta^{-1}(\hat{s})F^v(\hat{\mathbf{v}}, \hat{\mathbf{w}})\right)\mathbf{T}^{-1}(\hat{s}) + \text{h.o.t. in } \hat{\mathbf{w}}. \end{aligned} \tag{78}$$

Set $\hat{\mathbf{v}}^{(k)} = \mathbf{S}^{-1}\hat{\mathbf{u}}^{(k)}$ and $\hat{\mathbf{w}}^{(k)} = \hat{\mathbf{v}}^{(k)} - \hat{\mathbf{v}}$. Now

$$\begin{aligned} \mathbf{S}\hat{\mathbf{w}}^{(k)} &= \mathbf{T}^{-1}\left(\hat{s} + \hat{t}(\mathbf{w}^{(k-1)})\right)\left(\hat{\mathbf{v}} + \hat{\mathbf{w}}^{(k-1)}\right) - \mathbf{S}\hat{\mathbf{v}} \\ &= \mathbf{T}^{-1}(\hat{s})\left(\hat{\mathbf{v}} + \hat{\mathbf{w}}^{(k-1)}\right) - \mathbf{T}^{-1}(\hat{s})\mathbf{T}\left(\Delta^{-1}(\hat{s})F^v(\hat{\mathbf{v}}, \hat{\mathbf{w}}^{(k-1)})\right)\mathbf{T}^{-1}(\hat{s})\hat{\mathbf{v}} - \mathbf{S}\hat{\mathbf{v}} \\ &\quad + \text{h.o.t. in } \hat{\mathbf{w}}^{(k-1)} \\ &= \mathbf{T}^{-1}(\hat{s})\left(\hat{\mathbf{w}}^{(k-1)} - \mathbf{T}\left(\Delta^{-1}(\hat{s})F^v(\hat{\mathbf{v}}, \hat{\mathbf{w}}^{(k-1)})\right)\mathbf{T}^{-1}(\hat{s})\hat{\mathbf{v}}\right) + \text{h.o.t. in } \hat{\mathbf{w}}^{(k-1)} \\ &= \mathbf{T}^{-1}(\hat{s})\mathbf{R}(\hat{s}, \hat{\mathbf{v}})\hat{\mathbf{w}}^{(k-1)} + \text{h.o.t. in } \hat{\mathbf{w}}^{(k-1)}. \end{aligned} \tag{79}$$

Since $\mathbf{S}\hat{\mathbf{w}}^{(k)} = \hat{\mathbf{u}}^{(k)} - \hat{\mathbf{u}}_{\mathcal{A}}$ we have that

$$\hat{\mathbf{u}}^{(k)} - \hat{\mathbf{u}}_{\mathcal{A}} = \mathbf{T}^{-1}(\hat{s})\mathbf{R}(\hat{s}, \hat{\mathbf{v}})\mathbf{S}^{-1}\left(\hat{\mathbf{u}}^{(k-1)} - \hat{\mathbf{u}}_{\mathcal{A}}\right) + \text{h.o.t. in } \left(\hat{\mathbf{u}}^{(k-1)} - \hat{\mathbf{u}}_{\mathcal{A}}\right). \tag{80}$$

Equations (75) and (76) now follow from (80) and the fact that $\hat{\mu}^{(k)}$ is asymptotically given as a linear function of $\hat{\mathbf{u}}^{(k)}$. □

Adapting Theorem 3 to the context of Algorithm 1 we obtain the following Corollary.

Corollary 1 *Let $\mathbf{u}_{\mathcal{A}} \in W_{\mathcal{A}}^N$ be a fixed point of the Algorithm 1 and $(s, \mathbf{v}) \in W_{\mathcal{A}} \times W_{\mathcal{A}}^N$ a corresponding solution to (45). Let $\mu_{\mathcal{A}} \in W_{\mathcal{A}}$ be such that $P_{\mathcal{A}}(s\mu_{\mathcal{A}}) = 1$. Assume that $s_* := \inf_{y\in\Gamma} s(y) > 0$ and let $\psi_{\max}$ be as in Theorem 3. Then for any $\varepsilon > 0$ the iterates of Algorithm 1 satisfy*

$$||\mathbf{u}^{(k)} - \mathbf{u}_{\mathcal{A}}||_{L^2_v(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}} \leq \left(\frac{\psi_{\max}}{s_*} + \varepsilon\right) ||\mathbf{u}^{(k-1)} - \mathbf{u}_{\mathcal{A}}||_{L^2_v(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}}, \quad k \in \mathbb{N}$$

whenever $\mathbf{u}^{(k)}$ is sufficiently close to $\mathbf{u}_{\mathcal{A}}$. Furthermore, there exists $C > 0$ such that

$$||\mu^{(k)} - \mu_{\mathcal{A}}||_{L^2_v(\Gamma)} \leq C||\mathbf{u}^{(k)} - \mathbf{u}_{\mathcal{A}}||_{L^2_v(\Gamma)\otimes\mathbb{R}^N_{\mathbf{M}}}, \quad k \in \mathbb{N}. \tag{81}$$

Proof Interpret Theorem 3 in the context of Algorithm 1. The bound $\phi_{\min} \geq s_*$ is a consequence of Lemma 1. □

Obviously the previous Corollary has practical value only if $\psi_{\max} < s_*$. Here we will briefly discuss the value of $\psi_{\max}$ in the case that $(\hat{s}, \hat{\mathbf{v}}) \in \mathbb{R}^P \times \mathbb{R}^{PN}$ is associated to the dominant fixed point of Algorithm 2. Observe that the equation $\hat{\mathbf{z}} = \mathbf{R}(\hat{s}, \hat{\mathbf{v}})\hat{\mathbf{w}}$ is equivalent to the system

$$\begin{cases} \mathbf{z}(y) = \mathbf{w}(y) - P_{\mathcal{A}}(t\mathbf{u}_{\mathcal{A}})(y) \\ P_{\mathcal{A}}(st)(y) = P_{\mathcal{A}}\left(\langle \mathbf{v}, \mathbf{w}\rangle_{\mathbb{R}^N_{\mathbf{M}}}\right)(y) \end{cases} \tag{82}$$

for all $y \in \Gamma$. We see that, if $\mathbf{w} = \mathbf{v}$ then $\mathbf{z} = 0$, whereas, if $\langle \mathbf{w}(y), \mathbf{v}(y)\rangle_{\mathbb{R}^N_{\mathbf{M}}} = 0$ for all $y \in \Gamma$ then $\mathbf{z} = \mathbf{w}$. Thus, the matrix $\mathbf{R}(\hat{s}, \hat{\mathbf{v}})$ acts as a deflation that shrinks vectors that are close to $\mathbf{v}(y)$ and preserves vectors that are almost orthogonal to $\mathbf{v}(y)$. From Proposition 4 we know that $s^{-1}(y)$ is an approximation of the smallest eigenvalue of $\mathbf{M}^{-1}\mathbf{K}(y)$ and $\mathbf{v}(y)$ is an approximation of the corresponding eigenvector. By Lemma 1 the operator norm of $\mathbf{S}^{-1}$ is bounded by $\sup_{y\in\Gamma} \lambda_1^{-1}(y)$, where $\lambda_1(y)$ is the smallest eigenvalue of $\mathbf{M}^{-1}\mathbf{K}(y)$. Analogously, since the eigenvector corresponding to this smallest eigenvalue is deflated by $\mathbf{R}(\hat{s}, \hat{\mathbf{v}})$, we expect the norm of $\mathbf{R}(\hat{s}, \hat{\mathbf{v}})\mathbf{S}^{-1}$ to be bounded by a value close to $\sup_{y\in\Gamma} \lambda_2^{-1}(y)$, where $\lambda_2(y)$ is the second smallest eigenvalue of $\mathbf{M}^{-1}\mathbf{K}(y)$. With this reasoning, if the deflation is sufficient, there is $\psi^*_{\max} \in \mathbb{R}$ such that

$$\frac{\psi_{\max}}{s_*} \leq \frac{\psi^*_{\max}}{s_*} \approx \lambda_{1/2} := \frac{\sup_{y\in\Gamma} \lambda_2^{-1}(y)}{\inf_{y\in\Gamma} \lambda_1^{-1}(y)} = \frac{\sup_{y\in\Gamma} \lambda_1(y)}{\inf_{y\in\Gamma} \lambda_2(y)}. \tag{83}$$

One might suspect that the speed of convergence of the spectral inverse iteration is characterized by the largest value of the ratio $\lambda_1(y)/\lambda_2(y)$. The bound obtained from (83) is slightly more pessimistic, though not necessarily optimal.

5.2.4 Combined error analysis

Let $(\mu, u) \in L^2_\nu(\Gamma) \times L^2_\nu(\Gamma) \otimes H^1_0(D)$ be the smallest eigenvalue and the associated eigenfunction of the continuous problem (9). Let $(\mu_h, u_h) \in L^2_\nu(\Gamma) \times L^2_\nu(\Gamma) \otimes V_h$ be the corresponding eigenpair of the semi-discrete problem (13). Assume that there exists a dominant fixed point $\mathbf{u}_\mathcal{A} \in W^N_\mathcal{A}$ of Algorithm 1 and an associated eigenvalue approximation $\mu_{h,\mathcal{A}} := \mu_\mathcal{A} \in W_\mathcal{A}$ as in Proposition 4. Denote by $\mathbf{u}^{(k)} \in W^N_\mathcal{A}$ the k:th iterate of Algorithm 1 and by $\mu_{h,\mathcal{A},k} := \mu^{(k)} \in W_\mathcal{A}$ the associated solution to (37). Let $u_{h,\mathcal{A}}$ and $u_{h,\mathcal{A},k}$ denote the functions in $W_\mathcal{A} \otimes V_h$, whose coordinates are defined by the vectors $\mathbf{u}_\mathcal{A}$ and $\mathbf{u}^{(k)}$ respectively. The spatial, stochastic, and iteration errors may now be separated in the following sense:

$$\begin{aligned} ||\mu - \mu_{h,\mathcal{A},k}||_{L^2_\nu(\Gamma)} \leq\ & ||\mu - \mu_h||_{L^2_\nu(\Gamma)} \\ & + ||\mu_h - \mu_{h,\mathcal{A}}||_{L^2_\nu(\Gamma)} + ||\mu_{h,\mathcal{A}} - \mu_{h,\mathcal{A},k}||_{L^2_\nu(\Gamma)} \end{aligned} \tag{84}$$

and

$$\begin{aligned} ||u - u_{h,\mathcal{A},k}||_{L^2_\nu(\Gamma)\otimes L^2(D)} \leq\ & ||u - u_h||_{L^2_\nu(\Gamma)\otimes L^2(D)} + ||u_h - u_{h,\mathcal{A}}||_{L^2_\nu(\Gamma)\otimes L^2(D)} \\ & + ||u_{h,\mathcal{A}} - u_{h,\mathcal{A},k}||_{L^2_\nu(\Gamma)\otimes L^2(D)}. \end{aligned} \tag{85}$$

Under sufficient conditions we may now bound each term in the Eqs. (84) and (85) separately using the theory developed earlier in this section. The first term may be approximated using Theorem 1, the second term may be approximated using Theorem 2 and Proposition 2, and the third term may be approximated using Corollary 1 of Theorem 3 and the hypothesis (83). We therefore expect that, with an optimal choice the multi-index sets $\mathcal{A}_\epsilon$ for $\epsilon > 0$, the output of the spectral inverse iteration converges to the exact solution according to

$$||u - u_{h,\mathcal{A}_\epsilon,k}||_{L^2_\nu(\Gamma)\otimes L^2(D)} \lesssim h^{1+l} + (\#\mathcal{A}_\epsilon)^{-r} + \lambda_{1/2}^k \tag{86}$$

and similarly

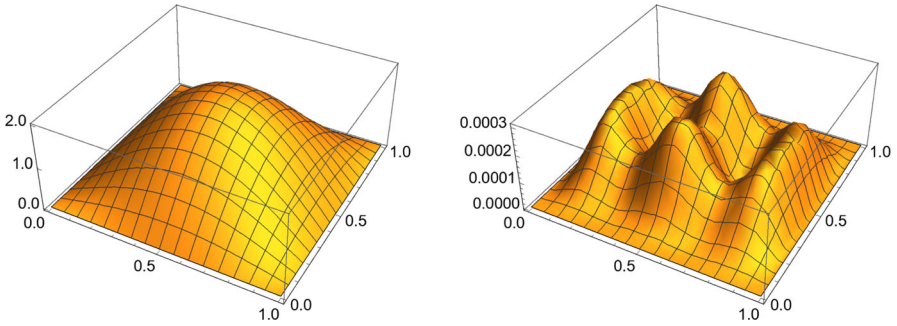
$$||\mu - \mu_{h,\mathcal{A}_\epsilon,k}||_{L^2_\nu(\Gamma)} \lesssim h^{2l} + (\#\mathcal{A}_\epsilon)^{-r} + \lambda_{1/2}^k \tag{87}$$

for certain rates $r > 0$ and $l > 0$.

5.3 Numerical examples

We present numerical evidence to verify the Eqs. (86) and (87). In each of the following examples we compute the smallest eigenvalue and the corresponding eigenfunction of the model problem (4) in the unit square $D = [0, 1]^2$ using the Algorithm 1. We use the smallest eigenvector at $y = 0$ as an initial guess. For the diffusion coefficient we assume the form (5) with $a_0 := 1$ and

$$a_m(x) := \begin{cases} (m+1)^{-\varsigma} \sin(m\pi x_1), & m = 1, 3, \dots \\ (m+1)^{-\varsigma} \sin(m\pi x_2), & m = 2, 4, \dots \end{cases} \quad x = (x_1, x_2) \in D,$$



(a) Mean $\mathbb{E}[u_*(x,\cdot)]$ **(b)** Variance $\mathrm{Var}[u_*(x,\cdot)]$

Fig. 1 The mean and variance of the eigenfunction as computed by Algorithm 1

where we set $\varsigma = 3.2$. Now $||a_m||_{L^\infty(D)} \leq Cm^{-\varsigma}$ and $||a_m||_{W^{2,\infty}(D)} \leq Cm^{-\varsigma+2}$ so that the assumptions (6)–(8) for $s = 2$ are satisfied with $p_0 > \varsigma^{-1}$ and $p_2 > (\varsigma-2)^{-1}$. We therefore expect the regions of analyticity in Proposition 1 to increase according to $\tau_m \geq Cm^{\varsigma-1}$.

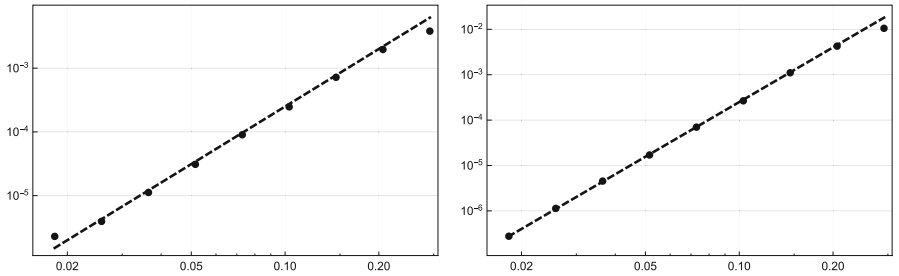
The deterministic mesh is a uniform grid of second order quadrilateral elements in all computations. The discretization in the parameter space is obtained by setting $\tau_m := (m+1)^{\varsigma-1}$ for $m = 1, 2, \ldots$ and using the multi-index sets $\mathcal{A}_\epsilon$ as defined in Proposition 2. Multi-index sets of this form have been introduced in [7] and in [5] an algorithm for generating them has been suggested.

We use a matrix free formulation of the conjugate gradient method for solving the linear systems (41) and (43). The preconditioner is constructed using the mean of the parametric matrix in question [17] and as an initial guess we set the solution of the system from the previous iteration. We wish to note that in this setting only a very few iterations of the conjugate gradient method are needed at each step of the spectral inverse iteration.

In the lack of an exact solution we compute an overkill solution (μ_*, u_*) for which the number of deterministic degrees of freedom is $N = 36741$, the parameter ϵ is chosen such that $\#\mathcal{A}_\epsilon = 264$, and the number of iterations is $k = 16$. This results in roughly 10^7 total degrees of freedom. The number of active dimensions in the overkill solution is $M(\mathcal{A}_\epsilon) = 113$. All the numerical examples in this section have been computed using this overkill solution as a reference. The expected value and variance of the eigenfunction are presented in Fig. 1.

5.3.1 Convergence in space

Keeping the number of stochastic degrees of freedom $\#\mathcal{A}_\epsilon = 264$ and the number of iterations $k = 16$ fixed, we may investigate the convergence of the solution $(\mu_{*,h}, u_{*,h})$ as a function of the spatial discretization parameter h. This convergence for piecewise quadratic basis functions is illustrated in Fig. 2. We observe algebraic convergence rates of order 3 and 4 for the eigenfunction and eigenvalue respectively, exactly as



(a) The errors $||u_* - u_{*,h}||_{L^2_\nu(\Gamma)\otimes L^2(D)}$. **(b)** The errors $||\mu_* - \mu_{*,h}||_{L^2_\nu(\Gamma)}$.

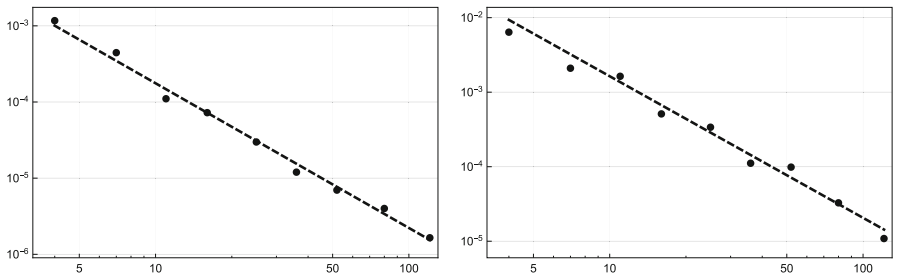
Fig. 2 Convergence of the spatial errors for the eigenfunction and eigenvalue as computed by Algorithm 1. The points represent a log–log plot of the errors as a function of h. The dashes lines represent the rates h^3 and h^4 respectively

predicted by Theorem 1. Thus, the error behaves like $N^{-3/2}$ and N^{-2} with respect to the number of deterministic degrees of freedom.

5.3.2 Convergence in the parameter domain

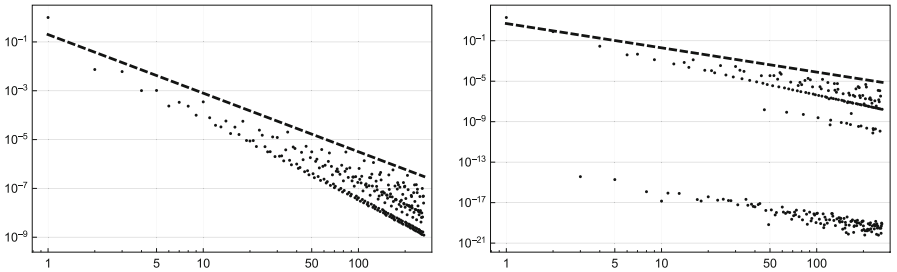
Keeping the number of spatial degrees of freedom $N = 36{,}741$ and the number of iterations $k = 16$ fixed, we may investigate the convergence of the solution $(\mu_{*,\mathcal{A}_\epsilon}, u_{*,\mathcal{A}_\epsilon})$ as a function of $\#\mathcal{A}_\epsilon$ as $\epsilon \to 0$. This convergence is illustrated in Fig. 3. We observe approximate algebraic convergence rates of order $-r = -1.9$ with respect to the number of stochastic degrees of freedom $\#\mathcal{A}_\epsilon$.

In Fig. 4 we have presented the norms of the Legendre coefficients of the overkill solution. The ordering of the coefficients is the same as the order in which they would appear in the multi-index set $\#\mathcal{A}_\epsilon$ as $\epsilon \to 0$. We see that the norms converge at the rate $-r - 1/2 = -2.4$ exactly as we would expect from the proof of Proposition 2. In Fig. 5 we have presented the norms of the same Legendre coefficients sorted by decreasing magnitude. From this Figure we estimate that, with an optimal selection of the multi-index sets we could in fact observe a rate of convergence $-r = -2.3$ for



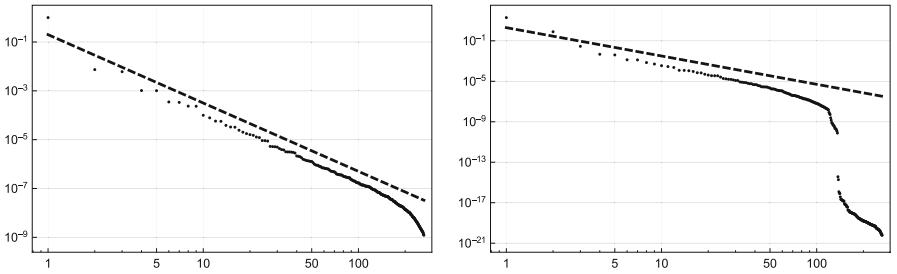
(a) The errors $||u_* - u_{*,\mathcal{A}_\epsilon}||_{L^2_\nu(\Gamma)\otimes L^2(D)}$. **(b)** The errors $||\mu_* - \mu_{*,\mathcal{A}_\epsilon}||_{L^2_\nu(\Gamma)}$.

Fig. 3 Convergence of the stochastic errors for the eigenfunction and eigenvalue as computed by Algorithm 1. The points represent a log–log plot of the errors as a function of $\#\mathcal{A}_\epsilon$. The dashed lines represent the rate $(\#\mathcal{A}_\epsilon)^{-1.9}$



(a) The values $||(u_*)_\alpha||_{L^2(D)}$. **(b)** The values $|(\mu_*)_\alpha|$.

Fig. 4 A log–log plot of the norms of the Legendre coefficients of the overkill solution. The dashed lines represent the algebraic rate -2.4



(a) The values $||(u_*)_\alpha||_{L^2(D)}$. **(b)** The values $|(\mu_*)_\alpha|$.

Fig. 5 A log–log plot of the norms of the Legendre coefficients of the overkill solution sorted by decrasing magnitude. The dashed lines represent the algebraic rate -2.8

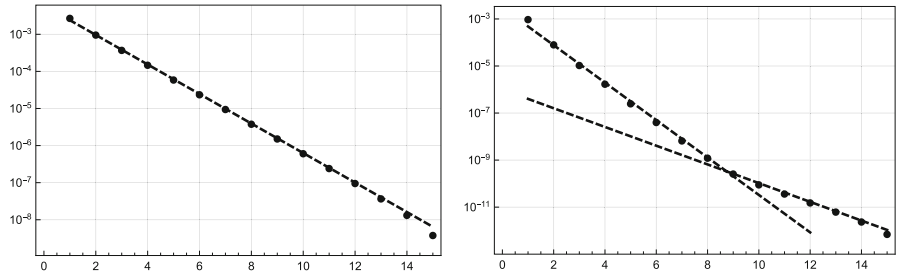
the error of the solution. This ideal rate of convergence is somewhat faster than the asymptotic theoretical bound of $-r = -\varsigma + 3/2 = -1.7$ predicted by Proposition 2.

Interestingly we observe two well separated clusters of values in Fig. 4b. It seems that many of the multi-indices that correspond to relatively large Legendre coefficients of the eigenfunction, account only for a marginal contribution to the eigenvalue.

5.3.3 Convergence of the iteration error

Keeping the number of spatial basis functions $N = 36{,}741$ and the parameter ϵ fixed so that $\#\mathcal{A}_\epsilon = 264$, we may investigate the convergence of the solution $(\mu_{*,k}, u_{*,k})$ as a function of the number of iterations k. This convergence is illustrated in Fig. 6. Assuming that the variation in the eigenvalues within the parameter space is small, the value $\lambda_{1/2}$ defined in (83) may be approximated by the ratio of the two smallest eigenvalues of the problem at $y = 0$. Thus, Fig. 6 suggests that the error behaves asymptotically like $\lambda_{1/2}^k$, just as predicted by Corollary 1.

It is worth noting that, from the analysis of the classical inverse iteration, one might expect the eigenvalue to converge faster than the eigenfunction. In fact, the eigenvalue exhibits a faster rate of convergence at first and the error behaves like $\lambda_{1/2}^{2k}$. Comparing to the results of the previous example, we see that $k \approx 9$ represents a turning point after which the stochastic error in the eigenfunction starts to dominate the iteration



(a) The errors $||u_* - u_{*,k}||_{L^2_\nu(\Gamma)\otimes L^2(D)}$. **(b)** The errors $||\mu_* - \mu_{*,k}||_{L^2_\nu(\Gamma)}$.

Fig. 6 Convergence of the iteration errors for the eigenfunction and eigenvalue as computed by Algorithm 1. The points represent a log plot of the errors as a function of k. The dashed lines represent the rates $\bar{\lambda}_{1/2}^{k}$ and $\bar{\lambda}_{1/2}^{2k}$, where $\bar{\lambda}_{1/2}$ is the ratio of the two smallest eigenvalues of the problem at $y = 0$

error. Hence, for $k \geq 9$ the polynomial approximation in the parameter domain is insufficient to guarantee the degree of accuracy that is required for the eigenvalue to exhibit the faster rate of convergence that is otherwise characteristic to it.

5.3.4 Concluding remarks and comparison to sparse collocation

Using the finest levels of discretization, i.e., $N = 9296$ degrees of freedom for approximation in space and $\#\mathcal{A}_\epsilon = 121$ degrees of freedom for approximation in the parameter domain, and computing $k = 9$ steps of the inverse iteration we obtain a solution for which the $L^2_\nu(\Gamma) \otimes L^2(D)$ error of the eigenfunction is approximately 3×10^{-6}. The number of total degrees of freedom in this case is more than 10^6 and the number of active dimensions is $M(\mathcal{A}_\epsilon) = 60$. The total computational time on a standard desktop machine is approximately five minutes, most of which is spent in the conjugate gradient method for the linear systems (41) and (43).

When the solution computed via the spectral inverse iteration is compared to the results of the non-composite version of the sparse collocation method introduced in [4] and employed in e.g. [1] (see equations (5.12)–(5.13) and (5.16)–(5.17)), the statistics of the two solutions seem to almost coincide. Again using the finest levels of discretization ($N = 9296$ and $\#\mathcal{A}_\epsilon = 121$) for both methods, the $L^2(D)$ errors of mean and variance of the eigenfunction are both less than 3×10^{-8} and the errors in the eigenvalue are less than 3×10^{-11} and 3×10^{-9} for the mean and variance respectively.

6 Spectral subspace iteration

In this section we extend the spectral inverse iteration to a spectral subspace iteration, with which we can compute dominant subspaces of the inverse of the parametric matrix under consideration. The underlying assumption is that the subspace is sufficiently smooth with respect to the parameters. Convergence of the spectral subspace iteration is verified through numerical experiments.

6.1 On the analyticity of finite dimensional subspaces

Let us consider invariant subspaces for which the corresponding cluster of eigenvalues is sufficiently well separated from the rest of the spectrum. Assume a cluster $\mathcal{M}(y) = \{\mu_q(y)\}_{q=1}^{Q}$ of eigenvalues of (9) so that

(i) each $\mu_q(y)$ is of finite multiplicity as an eigenvalue of $A(y)$ for all $y \in \Gamma$ and
(ii) the minimum spectral gap $\inf_{y\in\Gamma} \operatorname{dist}(\mathcal{M}(y), \sigma(A(y))\backslash\mathcal{M}(y))$ is positive.

It is in general difficult to consider the analyticity of each of the eigenmodes separately. However, we might still expect the associated invariant subspace to be analytic as a function of y. More precisely, let $\{u_q(y)\}_{q=1}^{Q'}$ be a maximal collection of linearly independent eigenfunctions corresponding to the eigenvalues $\mathcal{M}(y)$ for all $y \in \Gamma$. It is not completely unreasonable to assume that $\operatorname{span}\{u_q(y)\}_{q=1}^{Q'}$ is analytic, in a suitable sense, as a function of the parameter vector y. This assumption is the basis of our algorithm of spectral subspace iteration. For more information on the regularity of perturbed eigenvalues see [14,15].

6.2 Algorithm description

As with the classical subspace iteration, the idea in the spectral version is to perform inverse iteration for a set of vectors and orthogonalize these vectors at each step. Orthogonality should here be understood in a sense that the vectors are orthogonal for all points in the parameter space Γ. This can be approximately achieved by performing the Gram-Schmidt orthogonalization process for the vectors in the Galerkin sense, i.e., by projecting each elementary operation to the basis $W_{\mathcal{A}}$.

Fix a finite set of multi-indices $\mathcal{A} \subset (\mathbb{N}_0^\infty)_c$ and let $P = \#\mathcal{A}$. The spectral subspace iteration for the system (16) is now defined in Algorithm 3. Observe that, if the projections were precise, then the Algorithm would correspond to performing the classical subspace iteration pointwise on Γ. Orhtogonalization of the basis vectors via the Gram-Schmidt process is achieved in step (2). We expect Algorithm 3 to converge to an approximate basis for the Q-dimensional invariant subspace associated to the smallest eigenvalues of the system.

Algorithm 3 (Spectral subspace iteration) *Fix* $tol > 0$ *and let* $\{\mathbf{u}^{(0,q)}\}_{q=1}^{Q} \subset W_{\mathcal{A}}^N$ *be an initial guess for the basis of the subspace. For* $k = 1, 2, \ldots$ *do*

(1) *For each* $q = 1, \ldots, Q$ *solve* $\mathbf{v}^{(q)} \in W_{\mathcal{A}}^N$ *from the linear equation*

$$P_{\mathcal{A}}\left(\mathbf{K}\mathbf{v}^{(q)}\right) = \mathbf{M}\mathbf{u}^{(k-1,q)}. \tag{88}$$

(2) *For* $q = 1, \ldots, Q$ *do*

(2.1) *Set*

$$\mathbf{w}^{(q)} = \mathbf{v}^{(q)} - \sum_{i=1}^{q-1} P_{\mathcal{A}}\left(\mathbf{u}^{(k,i)} P_{\mathcal{A}}\left(\langle \mathbf{v}^{(q)}, \mathbf{u}^{(k,i)} \rangle_{\mathbb{R}^N_{\mathbf{M}}}\right)\right). \tag{89}$$

(2.2) *Solve* $s^{(q)} \in W_{\mathcal{A}}$ *from the nonlinear equation*

$$P_{\mathcal{A}}\left((s^{(q)})^2\right) = P_{\mathcal{A}}\left(||\mathbf{w}^{(q)}||^2_{\mathbb{R}^N_{\mathbf{M}}}\right). \tag{90}$$

(2.3) *Solve* $\mathbf{u}^{(k,q)} \in W^N_{\mathcal{A}}$ *from the linear equation*

$$P_{\mathcal{A}}\left(s^{(q)}\mathbf{u}^{(k,q)}\right) = \mathbf{w}^{(q)}. \tag{91}$$

(3) *Stop if a suitable criterion is satisfied and return* $\{\mathbf{u}^{(k,q)}\}_{q=1}^{Q} \subset W^N_{\mathcal{A}}$ *as the approximate basis for the subspace.*

In general we can not expect the output vectors $\{\mathbf{u}^{(k,q)}\}_{q=1}^{Q} \subset W^N_{\mathcal{A}}$ of Algorithm 3 to converge to any particular eigenvectors of the system (16). However, we still expect them to approximately span the subspace associated to the smallest eigenvalues of the system. In view of Sect. 6.1, if a cluster of eigenvalues is sufficiently well separated from the rest of the spectrum, then we assume the associated subspace to be analytic with respect to the parameter vector $y \in \Gamma$. In this case we may expect optimal convergence of the projections in the Algorithm.

Remark 5 In order to measure convergence of the Algorithm 3 we should be able to estimate the angle between subspaces over the parameter space Γ. It is not entirely trivial to perform this kind of a computation in practise. The numerical examples in Sect. 6.3 will hopefully give some more insight on this.

Remark 6 As noted in Sect. 3, the smallest eigenvalue of the problem (9) is always simple, hence analytic. For more general problems this might not be the case. For instance, in the event of an eigenvalue crossing, the eigenmode corresponding to the pointwise smallest eigenvalue is not (in general) even a continuous function of the parameter vector y. In this case we can modify the Algorithm 3 by adding the step

(2.0) Set $\mathbf{v}^{(1)} = \sum_{q=1}^{Q} \mathbf{v}^{(q)}$

before step (2.1). This should ensure optimal convergence, since even if the eigenmodes change places, we still expect their sum to be smooth with respect to y.

Using the tensors defined in Sect. 5 we may write Algorithm 3 in the following form.

Algorithm 4 (Spectral subspace iteration in tensor form) *Fix* $tol > 0$ *and let* $\{\hat{\mathbf{u}}^{(0,q)}\}_{q=1}^{Q} \subset \mathbb{R}^{PN}$ *be an initial guess for the basis of the subspace. For* $k = 1, 2, \ldots$ *do*

(1) *For each* $q = 1, \ldots, Q$ *solve* $\hat{\mathbf{v}}^{(q)} \in \mathbb{R}^{PN}$ *from the linear system*

$$\widehat{\mathbf{K}}\hat{\mathbf{v}}^{(q)} = \widehat{\mathbf{M}}\hat{\mathbf{u}}^{(k-1,q)}. \tag{92}$$

(2) *For* $q = 1, \ldots, Q$ *do*

(2.1) *Set*

$$\hat{\mathbf{w}}^{(q)} = \hat{\mathbf{v}}^{(q)} - \sum_{i=1}^{q-1} \mathbf{T}\left(F^v(\hat{\mathbf{v}}^{(q)}, \hat{\mathbf{u}}^{(k,i)})\right)\hat{\mathbf{u}}^{(k,i)}. \tag{93}$$

(2.2) *Solve* $\hat{s}^{(q)} \in \mathbb{R}^P$ *from the nonlinear system*

$$F(\hat{s}^{(q)}, \hat{\mathbf{w}}^{(q)}) = 0 \tag{94}$$

with the initial guess $s_\alpha^{(q)} = ||\hat{\mathbf{w}}^{(q)}||_{\mathbb{R}^P \otimes \mathbb{R}^N_{\mathbf{M}}} \delta_{\alpha 0}$ *for* $\alpha \in \mathcal{A}$.

(2.3) *Solve* $\hat{\mathbf{u}}^{(k,q)} \in \mathbb{R}^{PN}$ *from the linear system*

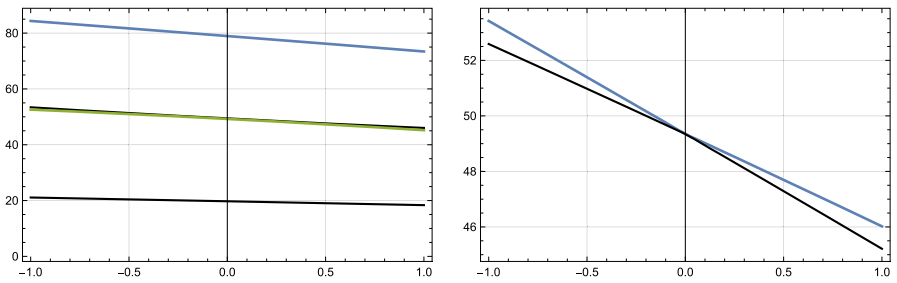
$$\mathbf{T}(\hat{s}^{(q)})\hat{\mathbf{u}}^{(k,q)} = \hat{\mathbf{w}}^{(q)}. \tag{95}$$

(3) *Stop if a suitable criterion is satisfied and return* $\{\hat{\mathbf{u}}^{(k,q)}\}_{q=1}^{Q} \subset \mathbb{R}^{PN}$ *as the approximate basis for the subspace.*

6.3 Numerical examples

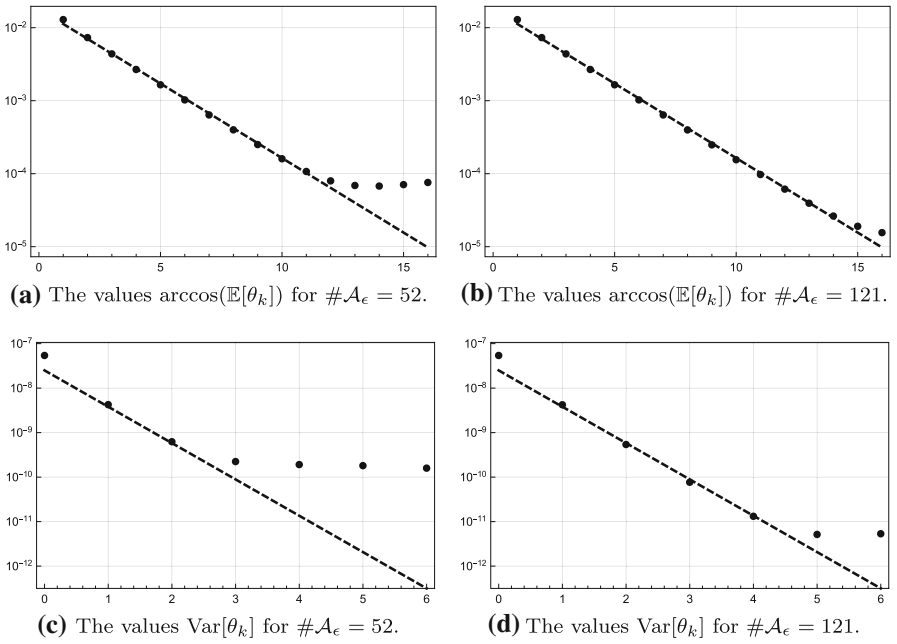
We use Algorithm 3 to compute the 3-dimensional subspace associated with the smallest eigenvalues of the model problem considered in Sect. 5.3. We let the deterministic mesh be a uniform grid of second order quadrilateral elements with $N = 2465$ degrees of freedom. As an initial guess we use the smallest eigenvectors of the problem at $y = 0$. In Fig. 7 we have presented the four smallest eigenvalues of the problem as a function of y_1, when $y_2, y_3, \ldots$ are held constant. We observe an eigenvalue crossing due to which the eigenvectors corresponding to the pointwise second and third smallest eigenvalues are discontinuous as functions of y.

In order to investigate the convergence of the spectral subspace iteration, we attempt to estimate the angle between the exact invariant subspace and the approximate one computed by Algorithm 3. For any fixed $y \in \Gamma$ we let $\mathbf{v}_1(y), \ldots, \mathbf{v}_Q(y)$ be a set of $\mathbb{R}^N_{\mathbf{M}}$ orthonormal exact eigenvectors corresponding to the Q-smallest eigenvalues of the problem. We define



(a) Eigenvalues 1-4. **(b)** Eigenvalues 2-3.

Fig. 7 A few smallest eigenvalues of the model problem as a function of y_1 when $y_2 = y_3 = \cdots = 0$. The smallest eigenvalue is well-separated. However, we observe a crossing of the second and third smallest eigenvalues



(a) The values $\arccos(\mathbb{E}[\theta_k])$ for $\#\mathcal{A}_\epsilon = 52$. **(b)** The values $\arccos(\mathbb{E}[\theta_k])$ for $\#\mathcal{A}_\epsilon = 121$.

(c) The values $\mathrm{Var}[\theta_k]$ for $\#\mathcal{A}_\epsilon = 52$. **(d)** The values $\mathrm{Var}[\theta_k]$ for $\#\mathcal{A}_\epsilon = 121$.

Fig. 8 Convergence of the Algorithm 3 for $Q = 3$. The points represent a log plot of approximate statistics of the error measure $\theta^{(k)}$ as a function of k. The dashed lines represent the rates $\bar{\lambda}_{3/4}^{k}$ and $\bar{\lambda}_{3/4}^{4k}$ for the top and bottom row plots respectively. Here $\bar{\lambda}_{3/4}$ is the ratio of the third and fourth smallest eigenvalues of the problem at $y = 0$

$$\theta_k(y) := |\det(\Theta^{(k)}(y))|,$$

where $\Theta^{(k)}(y) \in R^{Q\times Q}$ is a matrix with elements $\Theta_{ij}^{(k)}(y) = \langle \mathbf{u}_i^{(k)}(y), \mathbf{v}_j(y)\rangle_{\mathbb{R}_{\mathbf{M}}^N}$. Now $\theta_k(y)$ can be viewed as the cosine of the angle between the two subspaces at $y \in \Gamma$ (see for instance [10] formula (2.2)). Thus, convergence of the algorithm can be measured in terms of the statistics of θ_k. In the following examples we have estimated the mean and variance of θ_k using the non-composite version of the sparse collocation operator employed in [1]. For the definition of the collocation operator we have used the overkill multi-index set of Sect. 5.3 ($\#\mathcal{A}_\epsilon = 264$).

Convergence of the spectral subspace iteration for $Q = 3$ is illustrated in Fig. 8. We see that the values $\arccos(\mathbb{E}[\theta_k])$ behave like $\lambda_{3/4}^{k}$, where $\lambda_{3/4}$ is the ratio of the third and fourth smallest eigenvalues of the problem. Simultaneously the values $\mathrm{Var}[\theta_k]$ converge to zero. These results suggest that the angle between the exact subspace and the approximation computed by Algorithm 3 converges to zero on Γ. Furthermore, the rate of convergence is characterised by the rate $\lambda_{3/4}^{k}$ much like for the classical subspace iteration. Note however, that with a fixed basis for polynomial approximation, i.e. a fixed multi-index set $\mathcal{A}_\epsilon$, only a certain accuracy for the output may be reached. Increasing the number of basis polynomials makes more accurate solutions achievable.

7 Conclusions and future prospects

We have presented a comprehensive error analysis for the spectral inverse iteration, when applied to solving the ground state of a stochastic elliptic operator. We have also proposed a method of spectral subspace iteration and, using numerical examples, shown its potential in computing approximate subspaces associated to possibly clustered eigenvalues. Further analysis, both numerical and theoretical, of this algorithm is left for future research.

The numerical examples suggest that our algorithms are both accurate and efficient. However, theoretical estimates for the computational complexity are not entirely trivial to obtain as this would require information on the structure of the tensor of coeffiecients $c_{\alpha\beta\gamma}$. Moreover, when iterative solvers are used, the optimal strategy is to increase the associated tolerances in the course of the iteration. We note that sparse products of the spatial and stochastic approximation spaces, as in [5], may be applied to further reduce the computational effort, and that matrix free algorithms also allow for easy parallelization.

Acknowledgements Open access funding provided by Aalto University.

References

1. Andreev, R., Schwab, C.: Sparse tensor approximation of parametric eigenvalue problems. In: Lecture notes in computational science and engineering, vol. 83, pp. 203–241. Springer, Berlin (2012)
2. Babuška, I., Nobile, F., Tempone, R.: A stochastic collocation method for elliptic partial differential equations with random input data. SIAM J. Numer. Anal. **45**(3), 1005–1034 (2007)
3. Babuška, I., Osborn, J.: Eigenvalue problems. In: Handbook of Numerical Analysis, vol. II, pp. 641–787. Elsevier Science Publishers B.V., North-Holland (1991)
4. Bieri, M.: A sparse composite collocation finite element method for elliptic SPDEs. SIAM J. Numer. Anal. **49**(6), 2277–2301 (2011)
5. Bieri, M., Andreev, R., Schwab, C.: Sparse tensor discretization of elliptic SPDEs. SIAM J. Sci. Comput. **31**(6), 4281–4304 (2009)
6. Bieri, M., Andreev, R., Schwab, C.: Sparse tensor discretization of elliptic spdes. Tech. Rep. 2009-07, Seminar for Applied Mathematics, ETH Zürich, Switzerland. https://www.sam.math.ethz.ch/sam_reports/reports_final/reports2009/2009-07.pdf (2009)
7. Bieri, M., Schwab, C.: Sparse high order FEM for elliptic sPDEs. Comput. Methods Appl. Mech. Eng. **198**, 1149–1170 (2009)
8. Boffi, D.: Finite element approximation of eigenvalue problems. Acta Numer. **19**, 1–120 (2010)
9. Ghanem, R., Spanos, P.: Stochastic Finite Elements: A Spectral Approach. Dover Publications, Inc., Mineola (2003)
10. Gunawan, H., Neswan, O., Setya-Budhi, W.: A fromula for angles between subspaces of inner product spaces. Contrib. Algebra Geom. **46**(2), 311–320 (2005)
11. Hakula, H., Kaarnioja, V., Laaksonen, M.: Approximate methods for stochastic eigenvalue problems. Appl. Math. Comput. **267**(C), 664–681 (2015). https://doi.org/10.1016/j.amc.2014.12.112
12. Henrot, A.: Extremum Problems for Eigenvalues of Elliptic Operators. Birkhäuser, Basel (2006)
13. Kantorovich, L., Akilov, G.: Functional Analysis in Normed Spaces. Pergamon Press, New York (1964)
14. Kato, T.: Perturbation Theory for Linear Operators. Springer, Berlin (1997)

15. Kriegl, A., Michor, P., Rainer, A.: Denjoy-carleman differentiable perturbation of polynomials and unbounded operators. Integr. Equ. Oper. Theory **71**, 407–416 (2011)
16. Meidani, H., Ghanem, R.: Spectral power iterations for the random eigenvalue problem. AIAA J. **52**, 912–925 (2014)
17. Powell, C.E., Elman, H.C.: Block-diagonal preconditioning for spectral stochastic finite-element systems. IMA J. Numer. Anal. **29**(2), 350–375 (2008)
18. Soize, C., Ghanem, R.: Physical systems with random uncertainties: chaos representations with arbitrary probability measure. SIAM J. Sci. Comput. **26**, 395–410 (2004)
19. Sousedík, B., Elman, H.C.: Inverse subspace iteration for spectral stochastic finite element methods. SIAM/ASA J. Uncertain. Quantif. **4**, 163–189 (2016)
20. Verhoosel, C.V., Gutiérrez, M.A., Hulshoff, S.J.: Iterative solution of the random eigenvalue problem with application to spectral stochastic finite element systems. Int. J. Numer. Methods Eng. **68**, 401–424 (2006)
21. Xiu, D., Karniadakis, G.E.: The Wiener–Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. **24**, 619–644 (2002)