



Analysis of quasi-Monte Carlo methods for elliptic eigenvalue problems with stochastic coefficients

A. D. Gilbert¹ · I. G. Graham² · F. Y. Kuo³ · R. Scheichl^{1,2} · I. H. Sloan³

Received: 8 August 2018 / Revised: 20 March 2019 / Published online: 10 May 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

We consider the forward problem of uncertainty quantification for the generalised Dirichlet eigenvalue problem for a coercive second order partial differential operator with random coefficients, motivated by problems in structural mechanics, photonic crystals and neutron diffusion. The PDE coefficients are assumed to be uniformly bounded random fields, represented as infinite series parametrised by uniformly distributed i.i.d. random variables. The expectation of the fundamental eigenvalue of this problem is computed by (a) truncating the infinite series which define the coefficients; (b) approximating the resulting truncated problem using lowest order conforming finite elements and a sparse matrix eigenvalue solver; and (c) approximating the resulting finite (but high dimensional) integral by a randomly shifted quasi-Monte Carlo lattice rule, with specially chosen generating vector. We prove error estimates for the combined error, which depend on the truncation dimension s , the finite element mesh diameter h , and the number of quasi-Monte Carlo samples N . Under suitable regularity assumptions, our bounds are of the particular form $\mathcal{O}(h^2 + N^{-1+\delta})$, where $\delta > 0$ is arbitrary and the hidden constant is independent of the truncation dimension, which needs to grow as $h \rightarrow 0$ and $N \rightarrow \infty$. As for the analogous PDE source problem, the conditions under which our error bounds hold depend on a parameter $p \in (0, 1)$ representing the summability of the terms in the series expansions of the coefficients. Although the eigenvalue problem is nonlinear, which means it is generally considered harder than the source problem, in almost all cases ($p \neq 1$) we obtain error bounds that converge at the same rate as the corresponding rate for the source problem. The proof involves a detailed study of the regularity of the fundamental eigenvalue as a function of the random parameters. As a key intermediate result in the analysis, we prove that the spectral gap (between the fundamental and the second eigenvalues) is uniformly positive over all realisations of the random problem.

Mathematics Subject Classification 65D30 · 65N25 · 65N30

✉ A. D. Gilbert
a.gilbert@uni-heidelberg.de

Extended author information available on the last page of the article

1 Introduction

In this paper, we will propose methods for solving random 2nd-order elliptic eigenvalue problems (EVP) of the general form

$$-\nabla \cdot (a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y})) + b(\mathbf{x}, \mathbf{y}) u(\mathbf{x}, \mathbf{y}) = \lambda(\mathbf{y}) c(\mathbf{x}, \mathbf{y}) u(\mathbf{x}, \mathbf{y}), \quad \text{for } \mathbf{x} \in D, \quad (1.1)$$

where the derivative operator ∇ is with respect to the *physical variable* \mathbf{x} and where the *stochastic parameter*

$$\mathbf{y} = (y_j)_{j \in \mathbb{N}} \in U := [-\frac{1}{2}, \frac{1}{2}]^{\mathbb{N}}$$

is an infinite-dimensional vector of independently and identically distributed uniform random variables on $[-\frac{1}{2}, \frac{1}{2}]$. For simplicity, the physical domain $D \subset \mathbb{R}^d$, for $d = 1, 2, 3$, is assumed to be a bounded convex domain with Lipschitz boundary. To guarantee well-posedness of the eigenvalue problem (1.1), we impose homogeneous Dirichlet boundary conditions:

$$u(\mathbf{x}, \mathbf{y}) = 0 \quad \text{for } \mathbf{x} \in \partial D. \quad (1.2)$$

Under the initial assumption that $a(\cdot, \mathbf{y}), b(\cdot, \mathbf{y}), c(\cdot, \mathbf{y}) \in L^\infty(D)$, together with $a(\mathbf{x}, \mathbf{y}), c(\mathbf{x}, \mathbf{y}) \geq a_{\min} > 0$ for all $(\mathbf{x}, \mathbf{y}) \in D \times U$, the eigenvalues in (1.1) are real and bounded from below (this is a simple extension of the results in [38, Sec. 3.2]), and the leftmost (or *dominant*) eigenvalue λ_1 is simple. Since the coefficients depend on the stochastic parameters, the eigenvalues $\lambda(\mathbf{y})$ and corresponding eigenfunctions $u(\mathbf{x}, \mathbf{y})$ will also be stochastic.

Problems of the form (1.1) appear in many areas of engineering and physics. Two prominent examples are in nuclear reactor physics [14, 23, 33, 38] and in photonics [13, 19, 25, 28]. Problems of a similar type also appear in quantum physics, in acoustic, electromagnetic or elastic wave propagation, and in structural mechanics, where there is a huge engineering literature on the topic (see e.g. [18, 32]).

In nuclear reactor physics, the eigenproblem (1.1) corresponds to the mono-energetic diffusion approximation of the neutron transport equation [14, 23, 33, 38]. The dominant eigenvalue λ_1 of (1.1) describes the criticality of the reactor, while the corresponding eigenfunction u_1 models the associated neutron flux. The coefficient functions a , b and c correspond, respectively, to the diffusion coefficient, the absorption cross section and the fission cross section of the various materials in the reactor. These coefficients can vary strongly in \mathbf{x} and are subject to uncertainty in the composition of the constituent materials (e.g. liquid and vapour in the coolant), due to wear (e.g. “burnt” fuel) and due to geometric deviations from the original reactor design [3, 4, 39, 40].

In photonic band gap calculations in translationally invariant materials, e.g. photonic crystal fibres (PCFs), two decoupled eigenproblems of the type (1.1) have to be solved (with periodic boundary conditions): the transverse magnetic (TM) and the transverse electric (TE) mode problem [13, 19, 25, 28]. Here, $b \equiv 0$ and we have either

$a \equiv 1$ and $c = n^2$ (TM mode problem) or $a = 1/n^2$ and $c \equiv 1$ (TE mode problem), where $n = n(\mathbf{x}, \mathbf{y})$ is the refractive index of the PCF. The refractive index can be subject to uncertainty, due to heterogeneities or impurities in the material and due to geometric variations [15, 41].

The current paper demonstrates the power of quasi-Monte Carlo (QMC) methods for computing statistics of the eigenvalues of (1.1), (1.2). Our analysis will be restricted to approximating the expected value of the dominant eigenvalue λ_1 and linear functionals of the corresponding eigenfunction, but the method is applicable much more generally, and in particular to the applications listed above. We assume that, for all $\mathbf{x} \in D$ and $\mathbf{y} \in U$, the coefficients a and b admit series expansions of the following form:

$$a(\mathbf{x}, \mathbf{y}) = a_0(\mathbf{x}) + \sum_{j=1}^{\infty} y_j a_j(\mathbf{x}) \quad \text{and} \quad b(\mathbf{x}, \mathbf{y}) = b_0(\mathbf{x}) + \sum_{j=1}^{\infty} y_j b_j(\mathbf{x}), \quad (1.3)$$

and for the analysis assume that $c(\mathbf{x}, \mathbf{y}) = c(\mathbf{x})$. Although the fields a and b are parametrised by the same infinite sequence of random variables $(y_j)_{j \geq 0}$, this setting for the coefficients allows complete flexibility with respect to the correlation between a and b . To model two fields that are not correlated with each other, it suffices to set $a_{2j-1} \equiv 0$ and $b_{2j} \equiv 0$, for all $j \geq 1$. On the other hand, if there exists a $j \geq 1$ such that $a_j \not\equiv 0$ and $b_j \not\equiv 0$ then the two random fields will be correlated.

For a function $f : U \rightarrow \mathbb{R}$, its expected value with respect to the product uniform probability distribution is the infinite-dimensional integral:

$$\int_{[-\frac{1}{2}, \frac{1}{2}]^{\mathbb{N}}} f(\mathbf{y}) \, d\mathbf{y} := \lim_{s \rightarrow \infty} \int_{[-\frac{1}{2}, \frac{1}{2}]^s} f(y_1, \dots, y_s, 0, \dots) \, dy_1 \cdots dy_s,$$

provided that the limit exists. Our quantity of interest is then

$$\mathbb{E}_{\mathbf{y}} [\lambda_1] = \int_{[-\frac{1}{2}, \frac{1}{2}]^{\mathbb{N}}} \lambda_1(\mathbf{y}) \, d\mathbf{y}. \quad (1.4)$$

Our strategy for approximating (1.4) is to first truncate the expansions in (1.3) to s parameters by setting $y_j = 0$, for $j > s$, thus reducing (1.4) to a finite-dimensional quadrature problem. Then, for each $\mathbf{y} \in [-1/2, 1/2]^s$, we approximate (1.1) using a finite element (FE) discretisation on a mesh \mathcal{T}_h with mesh size h , to obtain a parametrised generalised matrix eigenproblem that can be solved iteratively (e.g. via an Arnoldi method or similar). The corresponding approximate dominant eigenvalue is denoted $\lambda_{1,s,h}(\mathbf{y})$. Now to approximate the integral in (1.4), we construct N suitable QMC quadrature points in the s -dimensional unit cube $[0, 1]^s$ via a rank-1 lattice rule with generating vector $\mathbf{z} \in \mathbb{N}^s$ (cf. [11]). The entire pointset is shifted by a uniformly-distributed random shift $\Delta \in [0, 1]^s$ and then translated into the cube $[-\frac{1}{2}, \frac{1}{2}]^s$. The final estimate of $\mathbb{E}_{\mathbf{y}} [\lambda_1]$ is then the (equal-weight) average of the approximate eigenvalues $\lambda_{1,s,h}$ at these N shifted QMC quadrature points and is denoted $Q_{N,s}(\mathbf{z}, \Delta)\lambda_{1,s,h}$.

The error depends on h , s and N and to estimate it we make some further assumptions on the coefficients, which are all detailed in Assumption A1. In particular, to bound the FE error (w.r.t. h), we require some spatial regularity of $(a_j)_{j \geq 0}$, $(b_j)_{j \geq 0}$ and c . To bound the dimension-truncation error (w.r.t. s) and the quadrature error (w.r.t. N), we assume p -summability of the sequences $(\|a_j\|_{L^\infty(D)})_{j \geq 0}$ and $(\|b_j\|_{L^\infty(D)})_{j \geq 0}$, for some $p \in (0, 1)$.

The main result in this paper is that, under these assumptions, there exists a constant independent of h , s and N such that

$$\sqrt{\mathbb{E}_\Delta \left[|\mathbb{E}_y[\lambda_1] - Q_{N,s}(z, \Delta)\lambda_{1,s,h}|^2 \right]} \leq C \left(h^2 + s^{-\frac{2}{p}+1} + N^{-\alpha} \right) \quad (1.5)$$

where $\alpha = \min(1-\delta, 1/p-1/2)$ for arbitrary $\delta \in (0, 1/2)$. This result, for which a full statement is given in Theorem 4.4 (along with a similar result for linear functionals \mathcal{G} of the corresponding eigenfunction), summarises the individual contributions to the overall error from the three approximations, i.e. discretisation (h), dimension-truncation (s) and quadrature (N). The errors in the three separate processes are established individually in Theorems 2.6, 4.1 and 4.2, respectively. To give a simple example of the power of estimate (1.5), if p is small enough, then the dimension-truncation error is negligible and the total error in (1.5) is bounded by the optimal FE convergence rate h^2 plus a QMC convergence rate that is arbitrarily close to N^{-1} . Importantly, the constant C does not depend on s .

A key result in obtaining (1.5) is Lemma 3.4, where we establish the regularity of λ_1 and u_1 with respect to y . The bounds on the mixed partial derivatives $|\partial_y^\nu \lambda_1(y)|$ of λ_1 are of product and order-dependent (POD) form, as in the case of (linear) boundary value problems [26]. Here, ν is a multi-index with finitely many non-zero entries $\nu_j \in \mathbb{N}$. The order dependence of the bounds in Lemma 3.4 is $(|\nu|!)^{1+\epsilon}$, for ϵ arbitrarily close to zero, which is only slightly larger than in the bounds in [26] and still allows us to achieve the (nearly) optimal dimension-independent QMC convergence rates. The constants in these bounds depend on the gap between $\lambda_1(y)$ and the second smallest eigenvalue $\lambda_2(y)$. Another important result is Proposition 2.4, where we prove that this gap is bounded away from zero uniformly in y under the assumptions which we shall make on a , b and c . Both Proposition 2.4 and Lemma 3.4 are essential components of our error analysis, however, they are also significant results in their own right.

Although stochastic eigenproblems have been of interest in engineering for some time, the mathematical literature is less developed. A common method of tackling these problems is the *reduced basis method* [16, 22, 27, 31], whereby the full parametric solution (eigenvalue) is approximated in a low-dimensional subspace that is constructed as the span of the solutions at specifically chosen parameter values. For the current work the most relevant paper is [1], where a sparse tensor approximation was used to estimate the expected value of the eigenvalue. A key result there is that simple eigenpairs are analytic with respect to the stochastic parameters, shown using the classical perturbation theory of Kato [24]. However, no bounds on the derivatives are given, which are required to theoretically justify the application of QMC rules. Here, we extend the results from [1] by proving explicit bounds on the derivatives, which in turn allows us to derive our *a priori* error bounds. Alternatively, this paper

can also be viewed as extending the results on QMC methods for stochastic elliptic source problems [26] to eigenvalue problems, and we remark that because of the nonlinearity of eigenvalue problems this is not merely a trivial extension. Despite the increased difficulty of this nonlinearity our error bound (1.5) achieves the same rates of convergence as the equivalent result for the PDE source problem in [26], for all $p \in (0, 1)$. The only difference is that our result does not hold for $p = 1$, in which case the result in [26] requires an additional assumption on the summability anyway.

The structure of the paper is as follows. In Sect. 2, we provide some relevant background theory of elliptic eigenproblems and of randomised lattice rules, and establish the FE error bound. Section 3 contains the parametric regularity analysis, which is then used in Sect. 4 to bound the quadrature and the truncation error. The paper concludes with two brief numerical experiments in Sect. 5, which demonstrate the sharpness of the bounds. An appendix contains the (technical) proof of the FE element error estimate.

2 Preliminary theory

In this section we present some preliminary theory on variational eigenvalue problems, FE discretisation and QMC methods. First we outline all of our assumptions on the coefficients, which, in particular, ensure that the problem (1.1) is well-posed.

- Assumption A1**
1. a and b are of the form (1.3) with $a_j, b_j \in L^\infty(D)$, for all $j \geq 0$, and $c \in L^\infty(D)$.
 2. There exists $a_{\min} > 0$ such that $a(\mathbf{x}, \mathbf{y}) \geq a_{\min}$, $b(\mathbf{x}, \mathbf{y}) \geq 0$ and $c(\mathbf{x}) \geq a_{\min}$, for all $\mathbf{x} \in D$, $\mathbf{y} \in U$.
 3. For some $p \in (0, 1)$,

$$\sum_{j=1}^{\infty} \|a_j\|_{L^\infty(D)}^p < \infty \quad \text{and} \quad \sum_{j=1}^{\infty} \|b_j\|_{L^\infty(D)}^p < \infty.$$

4. $a_j \in W^{1,\infty}(D)$ for $j \geq 0$ and

$$\sum_{j=1}^{\infty} \|a_j\|_{W^{1,\infty}(D)} < \infty.$$

The assumption that $b(\mathbf{x}, \mathbf{y}) \geq 0$ is made without loss of generality because any EVP with $b < 0$, but satisfying the rest of Assumption A1, can be shifted to an equivalent problem with “new b ” non-negative by adding $\sigma \cdot c(\mathbf{x}) \cdot u(\mathbf{x}, \mathbf{y})$ to both sides of (1.1), where σ is chosen such that $-b(\mathbf{x}, \mathbf{y}) \leq \sigma \cdot c(\mathbf{x})$ for all \mathbf{x}, \mathbf{y} . Such a σ exists due to Assumption A1. The eigenvalues of the original EVP are simply the eigenvalues of the shifted problem shifted by $-\sigma$, and the corresponding eigenspaces are unchanged. This has been used previously in, e.g., [19].

Throughout the paper, when it is unambiguous we will drop the \mathbf{x} -dependence when referring to a function defined on D at a parameter value \mathbf{y} . For example, we

will write the coefficients and eigenfunctions as $a(\mathbf{y}) := a(\cdot, \mathbf{y})$, $b(\mathbf{y}) := b(\cdot, \mathbf{y})$ and $u(\mathbf{y}) := u(\cdot, \mathbf{y})$.

Assumptions A1.1–A1.3 imply that for all $\mathbf{y} \in U$ we have $a(\mathbf{y}), b(\mathbf{y}) \in L^\infty(D)$. Furthermore, by the triangle inequality, the L^∞ -norms of these two coefficients can be bounded from above independently of \mathbf{y} : for all $\mathbf{x} \in D$, $\mathbf{y} \in U$

$$\|a(\mathbf{y})\|_{L^\infty(D)} \leq \|a_0\|_{L^\infty(D)} + \frac{1}{2} \sum_{j=1}^{\infty} \|a_j\|_{L^\infty(D)},$$

and similarly for $\|b(\mathbf{y})\|_{L^\infty(D)}$. For convenience, we define a single upper bound for all three coefficients

$$a_{\max} := \max \left\{ \sup_{\mathbf{y} \in U} \|a(\mathbf{y})\|_{L^\infty(D)}, \sup_{\mathbf{y} \in U} \|b(\mathbf{y})\|_{L^\infty(D)}, \|c\|_{L^\infty(D)} \right\}. \quad (2.1)$$

In Assumption A1.4, $W^{1,\infty}(D)$ is the usual Sobolev space of functions with essentially bounded gradient on D , to which we attach the norm $\|v\|_{W^{1,\infty}(D)} := \max\{\|v\|_{L^\infty(D)}, \|\nabla v\|_{L^\infty(D)}\}$. This assumption is needed to obtain the regularity result in Proposition 2.1.

2.1 Abstract theory for variational eigenproblems

To construct the variational formulation of the EVP (1.1), we introduce the test space $V := H_0^1(D)$, the usual first-order Sobolev space of real-valued functions with vanishing boundary trace, as well as its dual $V^* = H^{-1}(D)$, and equip V with the norm

$$\|v\|_V := \|\nabla v\|_{L^2(D)}.$$

We identify $L^2(D)$ with its dual and note that we have the following compact embeddings: $V \subset L^2(D) \subset V^*$. The $L^2(D)$ inner product is denoted by $\langle \cdot, \cdot \rangle$ and we use the same notation for the extension to the duality pairing on $V \times V^*$. Throughout the paper we shall repeatedly refer to the eigenvalues of the negative Laplacian on D , with boundary condition (1.2). These are strictly positive and, counting multiplicities, we denote them by

$$0 < \chi_1 < \chi_2 \leq \dots. \quad (2.2)$$

Multiplying each side of (1.1) by $v \in V$ and integrating (by parts) over D , we obtain the variational formulation

$$\begin{aligned} & \int_D a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x} + \int_D b(\mathbf{x}, \mathbf{y}) u(\mathbf{x}, \mathbf{y}) v(\mathbf{x}) \, d\mathbf{x} \\ &= \lambda(\mathbf{y}) \int_D c(\mathbf{x}) u(\mathbf{x}, \mathbf{y}) v(\mathbf{x}) \, d\mathbf{x} \quad \text{for all } v \in V. \end{aligned} \quad (2.3)$$

Correspondingly, for each \mathbf{y} we define the symmetric bilinear forms $\mathcal{A}(\mathbf{y}; \cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ by

$$\mathcal{A}(\mathbf{y}; w, v) := \int_D a(\mathbf{x}, \mathbf{y}) \nabla w(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x} + \int_D b(\mathbf{x}, \mathbf{y}) w(\mathbf{x}) v(\mathbf{x}) \, d\mathbf{x}, \quad (2.4)$$

and $\mathcal{M}(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ by

$$\mathcal{M}(w, v) := \int_D c(\mathbf{x}) w(\mathbf{x}) v(\mathbf{x}) \, d\mathbf{x}, \quad (2.5)$$

which are both inner products on their respective domains. Again, we will use the same notation for the corresponding duality pairings on $V \times V^*$. The norm induced by \mathcal{M} is denoted

$$\|v\|_{\mathcal{M}} = \sqrt{\mathcal{M}(v, v)}$$

and (using Assumption A1 and (2.1)) is equivalent to the $L^2(D)$ -norm:

$$\sqrt{a_{\min}} \|v\|_{L^2(D)} \leq \|v\|_{\mathcal{M}} \leq \sqrt{a_{\max}} \|v\|_{L^2(D)}, \quad \text{for } v \in L^2(D). \quad (2.6)$$

For each $\mathbf{y} \in U$, the variational eigenproblem equivalent to (1.1) is then: Find $0 \neq u(\mathbf{y}) \in V$ and $\lambda(\mathbf{y}) \in \mathbb{R}$ such that

$$\begin{aligned} \mathcal{A}(\mathbf{y}; u(\mathbf{y}), v) &= \lambda(\mathbf{y}) \mathcal{M}(u(\mathbf{y}), v), \quad \text{for all } v \in V, \\ \|u(\mathbf{y})\|_{\mathcal{M}} &= 1. \end{aligned} \quad (2.7)$$

In the following, let $\mathbf{y} \in U$ be fixed. The bilinear form $\mathcal{A}(\mathbf{y}; \cdot, \cdot)$ is coercive and bounded, uniformly in \mathbf{y} , i.e.,

$$\mathcal{A}(\mathbf{y}; v, v) \geq a_{\min} \|v\|_V^2, \quad \text{for all } v \in V, \quad \text{and} \quad (2.8)$$

$$\mathcal{A}(\mathbf{y}; w, v) \leq a_{\max} \left(1 + \chi_1^{-1}\right) \|w\|_V \|v\|_V, \quad \text{for all } w, v \in V, \quad (2.9)$$

with a_{\min} as in Assumption A1.2 and a_{\max} from (2.1), respectively. To establish (2.9) we have used the upper bound (2.1) and the Poincaré inequality:

$$\|v\|_{L^2(D)} \leq \chi_1^{-1/2} \|v\|_V, \quad \text{for } v \in V, \quad (2.10)$$

where for notational convenience we write the constant in terms of the Laplacian eigenvalue χ_1 , as defined above in (2.2). It is easy to see that the bound in (2.10) holds true and that equality is attained for $v = \phi_1$, the eigenfunction corresponding to χ_1 .

In addition to the variational form (2.7), it will also be convenient for us to study the corresponding solution operator, which we define now. Let $f \in V^*$ be arbitrary, and for each $\mathbf{y} \in U$, consider $T(\mathbf{y}) : V^* \rightarrow V$ given by

$$\mathcal{A}(\mathbf{y}; T(\mathbf{y})f, v) = \mathcal{M}(f, v) \quad \text{for all } v \in V. \quad (2.11)$$

Due to the symmetry of both $\mathcal{A}(\mathbf{y}; \cdot, \cdot)$ and $\mathcal{M}(\cdot, \cdot)$, each operator $T(\mathbf{y})$ is self-adjoint with respect to \mathcal{M} . Since $\mathcal{A}(\mathbf{y}; \cdot, \cdot)$ is coercive (2.8) and bounded (2.9), by the Lax–Milgram Theorem, see, e.g., [9], for every $f \in V^*$ there exists a unique solution $T(\mathbf{y})f \in V$ to (2.11), which satisfies $\|T(\mathbf{y})f\|_V \leq \|f\|_{V^*}/a_{\min}$. Hence, each $T(\mathbf{y}) : V^* \rightarrow V$ is bounded.

We can also consider the operators $T(\mathbf{y})$ from $L^2(D)$ to $L^2(D)$, in which case due to the compact embedding of V into $L^2(D)$, each $T(\mathbf{y}) : L^2(D) \rightarrow L^2(D)$ is compact. In this case, for $f \in L^2(D)$ the Lax–Milgram Theorem again gives a unique solution $T(\mathbf{y})f \in V$ with the bound

$$\|T(\mathbf{y})f\|_V \leq \frac{1}{\sqrt{\chi_1}} \frac{a_{\max}}{a_{\min}} \|f\|_{L^2(D)}, \quad (2.12)$$

where we have also used the equivalence of norms (2.6), along with the Cauchy–Schwarz and Poincaré inequalities to bound $\|f\|_{V^*}$.

From the spectral theory for compact, selfadjoint operators we know that each $T(\mathbf{y})$ has countably-many eigenvalues, which are all finite, real, strictly positive and have finite multiplicity (see, e.g., [9]). Counting multiplicities, the eigenvalues of $T(\mathbf{y})$ are denoted (in non-increasing order) by

$$\mu_1(\mathbf{y}) \geq \mu_2(\mathbf{y}) \geq \dots > 0,$$

with $\mu_k(\mathbf{y}) \rightarrow 0$ as $k \rightarrow \infty$. Comparing (2.7) with (2.11) we have that $\lambda(\mathbf{y})$ is an eigenvalue of (2.7) if and only if $\mu(\mathbf{y}) = 1/\lambda(\mathbf{y})$ is an eigenvalue of $T(\mathbf{y})$, and their eigenspaces coincide. It follows that (2.7) has countably-many eigenvalues $(\lambda_k(\mathbf{y}))_{k \in \mathbb{N}}$, which are all positive, have finite multiplicity and accumulate only at infinity. Counting multiplicities, we write them (in non-decreasing order) as

$$0 < \lambda_1(\mathbf{y}) \leq \lambda_2(\mathbf{y}) \leq \lambda_3(\mathbf{y}) \leq \dots .$$

with $\lambda_k(\mathbf{y}) \rightarrow \infty$ as $k \rightarrow \infty$. For an eigenvalue $\lambda(\mathbf{y})$ of (2.7) we define its eigenspace to be

$$E(\mathbf{y}, \lambda(\mathbf{y})) := \{u : u \text{ is an eigenfunction corresponding to } \lambda(\mathbf{y})\},$$

and from these eigenspaces, we can choose a sequence of eigenfunctions $(u_k(\mathbf{y}))_{k \in \mathbb{N}}$ corresponding to $(\lambda_k(\mathbf{y}))_{k \in \mathbb{N}}$ that form an orthonormal basis in V with respect to $\mathcal{M}(\cdot, \cdot)$. This again follows by the Spectral Theorem for $T(\mathbf{y})$.

The *min-max principle* (e.g. [6, (2.8)]) gives a representation of the k th eigenvalue as a minimum over all subspaces $S_k \subset V$ of dimension k :

$$\lambda_k(\mathbf{y}) = \min_{\substack{S_k \subset V \\ \dim(S_k)=k}} \max_{\substack{0 \neq u \in S_k}} \frac{\mathcal{A}(\mathbf{y}; u, u)}{\mathcal{M}(u, u)}. \quad (2.13)$$

We can use a combination of (2.1), (2.6) and (2.8) to bound both the numerator and denominator in the min-max representation (2.13) from above and below independently of \mathbf{y} , which in turn lets us bound the k th eigenvalue above and below independently of \mathbf{y} . Indeed, we have

$$\begin{aligned}\lambda_k(\mathbf{y}) &\geq \frac{a_{\min}}{a_{\max}} \min_{\substack{S_k \subset V \\ \dim(S_k)=k}} \max_{0 \neq u \in S_k} \frac{\langle \nabla u, \nabla u \rangle}{\langle u, u \rangle} \quad \text{and} \\ \lambda_k(\mathbf{y}) &\leq \frac{a_{\max}}{a_{\min}} \min_{\substack{S_k \subset V \\ \dim(S_k)=k}} \max_{0 \neq u \in S_k} \frac{\langle \nabla u, \nabla u \rangle + \langle u, u \rangle}{\langle u, u \rangle},\end{aligned}$$

but now the right hand side of both bounds contains the bilinear form corresponding to the negative Laplacian on D . Hence, using the min-max representation of the k th Laplacian eigenvalue χ_k the bounds on $\lambda_k(\mathbf{y})$ can be equivalently written as

$$\underline{\lambda}_k := \frac{a_{\min}}{a_{\max}} \chi_k \leq \lambda_k(\mathbf{y}) \leq \frac{a_{\max}}{a_{\min}} (\chi_k + 1) =: \overline{\lambda}_k. \quad (2.14)$$

Taking $v = u_k(\mathbf{y})$ as a test function in (2.7), we obtain

$$\lambda_k(\mathbf{y}) = \mathcal{A}(\mathbf{y}; u_k(\mathbf{y}), u_k(\mathbf{y})). \quad (2.15)$$

Then, using coercivity (2.8) as well as the upper bound (2.14) on $\lambda_k(\mathbf{y})$, we obtain

$$\|u_k(\mathbf{y})\|_V \leq \sqrt{\frac{\lambda_k(\mathbf{y})}{a_{\min}}} \leq \frac{\sqrt{a_{\max}(\chi_k + 1)}}{a_{\min}} =: \overline{u}_k. \quad (2.16)$$

Of particular interest is the smallest (also referred to as the minimal or fundamental) eigenvalue $\lambda_1(\mathbf{y})$. It follows by the Krein–Rutman Theorem that for every \mathbf{y} the fundamental eigenvalue $\lambda_1(\mathbf{y})$ is simple, see e.g. [20, Theorems 1.2.5 and 1.2.6]. The fact that $\lambda_1(\mathbf{y})$ is simple for all \mathbf{y} along with the uniform bound (2.14) ensures that the integrand (1.4) is well-defined. That the integrals $\mathbb{E}_{\mathbf{y}}[\lambda_1]$ and $\mathbb{E}_{\mathbf{y}}[\mathcal{G}(u_1)]$ (for $\mathcal{G} \in V^*$) exist follows by the truncation error bounds (4.1) and (4.2), respectively, which are shown later in Sect. 4.1 using Assumption A1.

In order to prove our finite element convergence results in Theorem 2.6, for $t \in [0, 1]$, we introduce the spaces:

$$Z^t := \left\{ v \in V : \Delta v \in H^{-1+t}(D) \right\}, \quad (2.17)$$

with norm

$$\|v\|_{Z^t} := \left(\|v\|_{L^2(D)}^2 + \|\Delta v\|_{H^{-1+t}(D)}^2 \right)^{\frac{1}{2}}. \quad (2.18)$$

where $H^r(D)$ is the usual fractional order Sobolev space (see [26]). When $t = 1$ we abbreviate Z^1 by Z . Since D is convex, $Z = V \cap H^{\frac{1}{2}}(D)$.

The following proposition shows that under Assumption A1, in particular A1.4, the eigenfunctions belong to $H^2(D)$ with norm bounded in terms of the corresponding eigenvalue.

Proposition 2.1 *Let $\mathbf{y} \in U$, let Assumption A1 hold and suppose $(\lambda(\mathbf{y}), u(\mathbf{y}))$ is an eigenpair of (2.7). Then $u(\mathbf{y}) \in Z = V \cap H^2(D)$ and there exists a constant $C > 0$ independent of \mathbf{y} , such that*

$$\|u(\mathbf{y})\|_Z \leq C\lambda(\mathbf{y}), \quad \text{for all } \mathbf{y} \in U. \quad (2.19)$$

Proof Since $u(\mathbf{y}) \in L^2(D)$ we can apply [26, Theorem 4.1] with $t = 1$ and $f = (\lambda(\mathbf{y})c - b(\mathbf{y}))u(\mathbf{y}) \in L^2(D)$ to give $u(\mathbf{y}) \in V \cap H^2(D)$ with

$$\begin{aligned} \|u(\mathbf{y})\|_Z &\leq C\|(\lambda(\mathbf{y})c - b(\mathbf{y}))u(\mathbf{y})\|_{L^2(D)} \\ &\leq C(\lambda(\mathbf{y})\|c\|_{L^\infty(D)} + \|b(\mathbf{y})\|_{L^\infty(D)})\|u(\mathbf{y})\|_{L^2(D)} \end{aligned}$$

Then, using (2.1), (2.6) and the normalisation in (2.7), we obtain

$$\begin{aligned} \|u(\mathbf{y})\|_Z &\leq C \frac{a_{\max}}{\sqrt{a_{\min}}} (\lambda(\mathbf{y}) + 1) \|u(\mathbf{y})\|_{\mathcal{M}} \\ &= C \frac{a_{\max}}{\sqrt{a_{\min}}} (\lambda(\mathbf{y}) + 1) \leq C \frac{a_{\max}}{\sqrt{a_{\min}}} \left(1 + \frac{1}{\lambda_1(\mathbf{y})}\right) \lambda(\mathbf{y}), \end{aligned}$$

and the result follows by the lower bound in (2.14). \square

2.2 Bounding the spectral gap

The Krein–Rutman theorem guarantees that the *spectral gap* $\lambda_2(\mathbf{y}) - \lambda_1(\mathbf{y})$ is positive for each parameter value \mathbf{y} . However, our estimates for the derivatives of $\lambda_1(\mathbf{y})$ proved in Sect. 3 require *uniform positivity* of this gap over all $\mathbf{y} \in U$. Here, we prove the required uniform positivity, using Assumption A1.3. We remark that this proof provides a justification for an assumption made previously without proof in [1].

The first step is the following elementary lemma, which shows that subsets of ℓ^∞ that are majorised by an ℓ^q sequence (for some $1 < q < \infty$) are compact.

Lemma 2.2 *Let $\boldsymbol{\alpha} \in \ell^q$ for some $1 < q < \infty$. The set $U(\boldsymbol{\alpha}) \subset \ell^\infty$ given by*

$$U(\boldsymbol{\alpha}) := \left\{ \mathbf{w} \in \ell^\infty : |w_j| \leq \frac{1}{2}|\alpha_j| \right\}$$

is a compact subset of ℓ^∞ .

Proof Since ℓ^∞ is a normed (and hence a metric) space, $U(\boldsymbol{\alpha})$ is compact if and only if it is sequentially compact. To show sequential compactness of $U(\boldsymbol{\alpha})$, take any sequence $\{\mathbf{y}^{(n)}\}_{n \geq 1} \subset U(\boldsymbol{\alpha})$. Clearly, by definition of $U(\boldsymbol{\alpha})$, each $\mathbf{y}^{(n)} \in \ell^q$ and moreover,

$$\|\mathbf{y}^{(n)}\|_{\ell^q} \leq \frac{1}{2}\|\boldsymbol{\alpha}\|_{\ell^q} < \infty \quad \text{for all } n \in \mathbb{N}.$$

So $\mathbf{y}^{(n)}$ is a bounded sequence in ℓ^q . Since $q < \infty$, ℓ^q is a reflexive Banach space, and so by [9, Theorem 3.18] $\{\mathbf{y}^{(n)}\}_{n \geq 1}$ has a subsequence that converges weakly to a limit in ℓ^q . We denote this limit by \mathbf{y}^* , and, with a slight abuse of notation, we denote the convergent subsequence again by $\{\mathbf{y}^{(n)}\}_{n \geq 1}$.

We now prove that $\mathbf{y}^* \in U(\boldsymbol{\alpha})$ and that the weak convergence is in fact strong, i.e. we show $\mathbf{y}^{(n)} \rightarrow \mathbf{y}^*$ in ℓ^∞ , as $n \rightarrow \infty$. For any $j \in \mathbb{N}$, consider the linear functional $f_j : \ell^q \rightarrow \mathbb{R}$ given by $f_j(\mathbf{w}) = w_j$, where w_j denotes the j th element of the sequence $\mathbf{w} = (w_j)_{j \geq 1} \in \ell^q$. Clearly, $f_j \in (\ell^q)^*$ (the dual space) and using the weak convergence established above, it follows that

$$y_j^{(n)} = f_j(\mathbf{y}^{(n)}) \rightarrow f_j(\mathbf{y}^*) = y_j^* \text{ as } n \rightarrow \infty, \quad \text{for each fixed } j.$$

That is, we have componentwise convergence. Furthermore, since $|y_j^{(n)}| \leq \frac{1}{2}|\alpha_j|$ it follows that $|y_j^*| \leq \frac{1}{2}|\alpha_j|$ for each j , and hence $\mathbf{y}^* \in U(\boldsymbol{\alpha})$.

Now, for any $J \in \mathbb{N}$ we can write

$$\begin{aligned} \|\mathbf{y}^{(n)} - \mathbf{y}^*\|_{\ell^q}^q &= \sum_{j=1}^J |y_j^{(n)} - y_j^*|^q + \sum_{j=J+1}^{\infty} |y_j^{(n)} - y_j^*|^q \\ &\leq J \max_{j=1,2,\dots,J} |y_j^{(n)} - y_j^*|^q + \sum_{j=J+1}^{\infty} |\alpha_j|^q. \end{aligned} \quad (2.20)$$

Let $\varepsilon > 0$. Since $\boldsymbol{\alpha} \in \ell^q$, we can choose $J \in \mathbb{N}$ such that

$$\sum_{j=J+1}^{\infty} |\alpha_j|^q \leq \frac{\varepsilon^q}{2},$$

and since $\mathbf{y}^{(n)}$ converges componentwise we can choose $K \in \mathbb{N}$ such that

$$|y_j^{(n)} - y_j^*| \leq (2J)^{-1/q} \varepsilon \quad \text{for all } j = 1, 2, \dots, J \text{ and } n \geq K.$$

Thus, by (2.20) we have that $\|\mathbf{y}^{(n)} - \mathbf{y}^*\|_{\ell^q}^q \leq \varepsilon^q$ for all $n \geq K$, and hence that $\|\mathbf{y}^{(n)} - \mathbf{y}^*\|_{\ell^q} \rightarrow 0$ as $n \rightarrow \infty$.

Because $\|\mathbf{w}\|_{\ell^\infty} \leq \|\mathbf{w}\|_{\ell^q}$ when $\mathbf{w} \in \ell^q$ and $1 < q < \infty$, this also implies that $\mathbf{y}^{(n)} \rightarrow \mathbf{y}^*$ in ℓ^∞ , completing the proof. \square

A key property following from the perturbation theory of Kato [24] is that the eigenvalues $\lambda_k(\mathbf{y})$ are continuous in \mathbf{y} , which for completeness is shown below in Proposition 2.3. First, recall that $T(\mathbf{y})$ is the solution operator as defined in (2.11), and let $\Sigma(T(\mathbf{y}))$ denote the spectrum of $T(\mathbf{y})$.

Proposition 2.3 *Let Assumption A1 hold. Then the eigenvalues $\lambda_1, \lambda_2, \dots$ are Lipschitz continuous in \mathbf{y} .*

Proof We prove the result by establishing the continuity of the eigenvalues $\mu_k(\mathbf{y})$ of $T(\mathbf{y})$. Let $\mathbf{y}, \mathbf{y}' \in U$ and consider the operators $T(\mathbf{y}), T(\mathbf{y}') : L^2(D) \rightarrow L^2(D)$ as defined in (2.11). Since $T(\mathbf{y}), T(\mathbf{y}')$ are bounded and self-adjoint with respect to \mathcal{M} , it follows from [24, V, §4.3 and Theorem 4.10] that we have the following notion of continuity of $\mu(\cdot)$ in terms of $T(\cdot)$

$$\sup_{\mu \in \Sigma(T(\mathbf{y}))} \text{dist}(\mu, \Sigma(T(\mathbf{y}'))) \leq \|T(\mathbf{y}) - T(\mathbf{y}')\|_{L^2(D) \rightarrow L^2(D)}. \quad (2.21)$$

For an eigenvalue $\mu_k(\mathbf{y}) \in \Sigma(T(\mathbf{y}))$, (2.21) implies that there exists a $\mu_{k'}(\mathbf{y}') \in \Sigma(T(\mathbf{y}'))$ such that

$$|\mu_k(\mathbf{y}) - \mu_{k'}(\mathbf{y}')| \leq \|T(\mathbf{y}) - T(\mathbf{y}')\|_{L^2(D) \rightarrow L^2(D)}. \quad (2.22)$$

Note that this means there exists an eigenvalue of $T(\mathbf{y}')$ close to $\mu_k(\mathbf{y})$, but does not imply that the k th eigenvalue of $T(\mathbf{y}')$ is close to $\mu_k(\mathbf{y})$, that is, in (2.22) k is not necessarily equal to k' . However, consider any $\mu_k(\mathbf{y})$ and let m denote its multiplicity. Since $m < \infty$, we can assume without loss of generality that the collection $\mu_k(\mathbf{y}) = \mu_{k+1}(\mathbf{y}) = \dots = \mu_{k+m-1}(\mathbf{y})$ is a *finite system* of eigenvalues in the sense of Kato. It then follows from the discussion in [24, IV, §3.5] that the eigenvalues in this system depend continuously on the operator with multiplicity preserved. This preservation of multiplicity is key to our argument, since it states that for $T(\mathbf{y}')$ sufficiently close to $T(\mathbf{y})$ there are m consecutive eigenvalues $\mu_{k'}(\mathbf{y}'), \mu_{k'+1}(\mathbf{y}'), \dots, \mu_{k'+m-1}(\mathbf{y}') \in \Sigma(T(\mathbf{y}'))$, no longer necessarily equal, that are close to $\mu_k(\mathbf{y})$.

A simple argument then shows that each μ_k is continuous in the following sense

$$|\mu_k(\mathbf{y}) - \mu_k(\mathbf{y}')| \leq \|T(\mathbf{y}) - T(\mathbf{y}')\|_{L^2(D) \rightarrow L^2(D)}. \quad (2.23)$$

To see this, consider, for $k = 1, 2, \dots$, the graphs of μ_k on U . Note that the separate graphs can touch (and in principle can even coincide over some subset of U), but by definition cannot cross (since at every point in U the successive eigenvalues are nonincreasing); and by the preservation of multiplicity a graph cannot terminate and a finite set of graphs cannot change multiplicity at an interior point. Thus by (2.23) the ordered eigenvalues μ_k must be continuous for each $k \geq 1$ and satisfy (2.23).

It then follows from the relationship $\mu_k(\mathbf{y}) = 1/\lambda_k(\mathbf{y})$ along with the upper bound in (2.14) that we have a similar result for the eigenvalues λ_k of (2.7):

$$|\lambda_k(\mathbf{y}) - \lambda_k(\mathbf{y}')| \leq \overline{\lambda}_k^{-2} \|T(\mathbf{y}) - T(\mathbf{y}')\|_{L^2(D) \rightarrow L^2(D)}. \quad (2.24)$$

All that remains is to bound the right hand of (2.24) by $C_{\text{Lip}} \|\mathbf{y} - \mathbf{y}'\|_{\ell^\infty}$, with $C_{\text{Lip}} > 0$ independent of \mathbf{y} and \mathbf{y}' . To this end, note that since the right hand side of (2.11) is independent of \mathbf{y} we have

$$\mathcal{A}(\mathbf{y}; T(\mathbf{y})f, v) = \mathcal{A}(\mathbf{y}'; T(\mathbf{y}')f, v) \quad \text{for all } f \in L^2(D), v \in V.$$

Rearranging and then expanding this gives

$$\begin{aligned} \mathcal{A}(\mathbf{y}; (T(\mathbf{y}) - T(\mathbf{y}')) f, v) &= \mathcal{A}(\mathbf{y}'; T(\mathbf{y}') f, v) - \mathcal{A}(\mathbf{y}; T(\mathbf{y}') f, v) \\ &= \int_D \left([a(\mathbf{x}, \mathbf{y}') - a(\mathbf{x}, \mathbf{y})] \nabla[T(\mathbf{y}') f](\mathbf{x}) \cdot \nabla v(\mathbf{x}) \right. \\ &\quad \left. + [b(\mathbf{x}, \mathbf{y}') - b(\mathbf{x}, \mathbf{y})] [T(\mathbf{y}') f](\mathbf{x}) v(\mathbf{x}) \right) d\mathbf{x}. \end{aligned}$$

Letting $v = (T(\mathbf{y}) - T(\mathbf{y}'))f \in V$, the left hand side can be bounded from below using the coercivity (2.8) of $\mathcal{A}(\mathbf{y}; \cdot, \cdot)$, and the right hand side can be bounded from above using the Cauchy–Schwarz inequality to give

$$\begin{aligned} a_{\min} \| (T(\mathbf{y}) - T(\mathbf{y}')) f \|_V^2 &\leq \max \left(\|a(\mathbf{y}) - a(\mathbf{y}')\|_{L^\infty(D)}, \|b(\mathbf{y}) - b(\mathbf{y}')\|_{L^\infty(D)} \right) \\ &\quad \cdot \left(\|T(\mathbf{y}') f\|_V \| (T(\mathbf{y}) - T(\mathbf{y}')) f \|_V \right. \\ &\quad \left. + \|T(\mathbf{y}') f\|_{L^2(D)} \| (T(\mathbf{y}) - T(\mathbf{y}')) f \|_{L^2(D)} \right). \end{aligned}$$

Applying the Poincaré inequality (2.10) to both L^2 -norm factors, dividing by $a_{\min} \| (T(\mathbf{y}) - T(\mathbf{y}')) f \|_V$ and using the upper bound in (2.12) we have

$$\begin{aligned} &\| (T(\mathbf{y}) - T(\mathbf{y}')) f \|_V \\ &\leq \frac{a_{\max}(1 + 1/\chi_1)}{a_{\min}^2 \sqrt{\chi_1}} \|f\|_{L^2(D)} \max \left(\|a(\mathbf{y}) - a(\mathbf{y}')\|_{L^\infty(D)}, \|b(\mathbf{y}) - b(\mathbf{y}')\|_{L^\infty(D)} \right). \end{aligned}$$

Then, applying the Poincaré inequality (2.10) to the left hand side and taking the supremum over $f \in L^2(D)$ with $\|f\|_{L^2(D)} \leq 1$, in the operator norm we have

$$\begin{aligned} &\|T(\mathbf{y}) - T(\mathbf{y}')\|_{L^2(D) \rightarrow L^2(D)} \\ &\leq \frac{a_{\max}(\chi_1 + 1)}{a_{\min}^2 \chi_1^2} \max \left(\|a(\mathbf{y}) - a(\mathbf{y}')\|_{L^\infty(D)}, \|b(\mathbf{y}) - b(\mathbf{y}')\|_{L^\infty(D)} \right). \end{aligned}$$

Using this inequality as an upper bound for (2.24) we have that the eigenvalues inherit the continuity of the coefficients

$$|\lambda_k(\mathbf{y}) - \lambda_k(\mathbf{y}')| \leq C \max \left(\|a(\mathbf{y}) - a(\mathbf{y}')\|_{L^\infty(D)}, \|b(\mathbf{y}) - b(\mathbf{y}')\|_{L^\infty(D)} \right), \quad (2.25)$$

where

$$C = \overline{\lambda}_k^{-2} \frac{a_{\max}(\chi_1 + 1)}{a_{\min}^2 \chi_1^2} < \infty,$$

is clearly independent of \mathbf{y} and \mathbf{y}' .

Finally, to establish Lipschitz continuity with respect to \mathbf{y} , we recall Assumptions A1.1 and A1.3, expand the coefficients in (2.25) above and use the triangle inequality to give

$$\begin{aligned} |\lambda_k(\mathbf{y}) - \lambda_k(\mathbf{y}')| &\leq C \sum_{j=1}^{\infty} |y_j - y'_j| \max(\|a_j\|_{L^\infty(D)}, \|b_j\|_{L^\infty(D)}) \\ &\leq C \left(\sum_{j=1}^{\infty} \max(\|a_j\|_{L^\infty(D)}, \|b_j\|_{L^\infty(D)}) \right) \|\mathbf{y} - \mathbf{y}'\|_{\ell^\infty}. \end{aligned}$$

By Assumption A1 the sum is finite, and hence the eigenvalue $\lambda_k(\mathbf{y})$ is Lipschitz in \mathbf{y} . \square

Now that we have shown Lipschitz continuity of the eigenvalues and identified suitable compact subsets, we can show that the spectral gap is bounded uniformly away from 0.

Proposition 2.4 *Let Assumption A1 hold. Then there exists a $\delta > 0$, independent of \mathbf{y} , such that*

$$\lambda_2(\mathbf{y}) - \lambda_1(\mathbf{y}) \geq \delta. \quad (2.26)$$

Proof The idea of the proof is to rewrite $a(\mathbf{x}, \mathbf{y})$ as

$$a(\mathbf{x}, \mathbf{y}) = a_0(\mathbf{x}) + \sum_{j=1}^{\infty} \tilde{y}_j \tilde{a}_j(\mathbf{x}),$$

with $\tilde{y}_j = \alpha_j y_j$ and $\tilde{a}_j(\mathbf{x}) = a_j(\mathbf{x})/\alpha_j$. We then choose $\boldsymbol{\alpha} \in \ell^q$ to decay slowly enough so that $\sum_{j=1}^{\infty} \|\tilde{a}_j\|_{L^\infty(D)} < \infty$ and we apply a similar reparametrisation procedure to $b(\mathbf{x}, \mathbf{y})$. Then using the intermediate result (2.25) from the proof of Proposition 2.3 we can show that the eigenvalues of the “reparametrised” problem are continuous in the new parameter $\tilde{\mathbf{y}}$, which now ranges over the compact set $U(\boldsymbol{\alpha})$. The required bound on the spectral gap is obtained by using the equivalence of the eigenvalues of the original and reparametrised problems.

To give some details, note that there is no loss of generality in assuming $p > 1/2$, because if Assumption A1.3 holds with exponent $p' \leq 1/2$ then it also holds for all $p \in (p', 1)$. We consequently set $\varepsilon = 1 - p \in (0, 1/2)$ and consider the sequence $\boldsymbol{\alpha}$ defined by

$$\alpha_j = \|a_j\|_{L^\infty(D)}^\varepsilon + \|b_j\|_{L^\infty(D)}^\varepsilon + 1/j, \quad \text{for each } j = 1, 2, \dots \quad (2.27)$$

Setting $q = p/\varepsilon = p/(1-p) \in (1, \infty)$, using Assumption A1.3 and the triangle inequality, it is easy to see that $\boldsymbol{\alpha} \in \ell^q$. Moreover, the inclusion of $1/j$ in (2.27) ensures that $\alpha_j \neq 0$, for all $j \geq 1$. Hence, from now on, for $\mathbf{w} = (w_j)_{j=1}^{\infty} \in \ell^\infty$, we

can define the sequences $\boldsymbol{\alpha} \mathbf{w} = (\alpha_j w_j)_{j=1}^\infty$ and $\mathbf{w}/\boldsymbol{\alpha} = (w_j/\alpha_j)_{j=1}^\infty$. Then, recalling the definition of $U(\boldsymbol{\alpha})$ in Lemma 2.2, it is easy to see that

$\tilde{\mathbf{y}} \in U(\boldsymbol{\alpha})$ if and only if $\tilde{\mathbf{y}}/\boldsymbol{\alpha} \in U$ and moreover $\mathbf{y} \in U$ if and only if $\boldsymbol{\alpha} \mathbf{y} \in U(\boldsymbol{\alpha})$.

Now for $\mathbf{x} \in D$ and $\tilde{\mathbf{y}} \in U(\boldsymbol{\alpha})$, we define

$$\tilde{a}(\mathbf{x}, \tilde{\mathbf{y}}) = a_0(\mathbf{x}) + \sum_{j=1}^{\infty} \tilde{y}_j \frac{a_j(\mathbf{x})}{\alpha_j} \quad \text{and} \quad \tilde{b}(\mathbf{x}, \tilde{\mathbf{y}}) = b_0(\mathbf{x}) + \sum_{j=1}^{\infty} \tilde{y}_j \frac{b_j(\mathbf{x})}{\alpha_j},$$

from which it is easily seen that

$$\tilde{a}(\mathbf{x}, \tilde{\mathbf{y}}) = a(\mathbf{x}, \tilde{\mathbf{y}}/\boldsymbol{\alpha}) \quad \text{and} \quad \tilde{b}(\mathbf{x}, \tilde{\mathbf{y}}) = b(\mathbf{x}, \tilde{\mathbf{y}}/\boldsymbol{\alpha}). \quad (2.28)$$

Then we set

$$\tilde{\mathcal{A}}(\tilde{\mathbf{y}}; w, v) := \int_D (\tilde{a}(\mathbf{x}, \tilde{\mathbf{y}}) \nabla w(\mathbf{x}) \cdot \nabla v(\mathbf{x}) + \tilde{b}(\mathbf{x}, \tilde{\mathbf{y}}) w(\mathbf{x}) v(\mathbf{x})) \, d\mathbf{x} \quad \text{for } w, v \in V,$$

and we consider the reparametrised eigenvalue problem find $\tilde{\lambda}(\tilde{\mathbf{y}}) \in \mathbb{R}$ and $0 \neq \tilde{u}(\tilde{\mathbf{y}}) \in V$ such that

$$\begin{aligned} \tilde{\mathcal{A}}(\tilde{\mathbf{y}}; \tilde{u}(\tilde{\mathbf{y}}), v) &= \tilde{\lambda}_k(\tilde{\mathbf{y}}) \mathcal{M}(\tilde{u}(\tilde{\mathbf{y}}), v) \quad \text{for all } v \in V, \\ \|\tilde{u}(\tilde{\mathbf{y}})\|_{\mathcal{M}} &= 1. \end{aligned} \quad (2.29)$$

Note that because we have equality between the original and reparametrised coefficients (2.28), for each $\mathbf{y} \in U$, and corresponding $\tilde{\mathbf{y}} = \boldsymbol{\alpha} \mathbf{y} \in U(\boldsymbol{\alpha})$, (2.28) implies that there is equality between eigenvalues $\lambda_k(\mathbf{y})$ of (2.7) and $\tilde{\lambda}_k(\tilde{\mathbf{y}})$ of the reparametrised eigenvalue problem (2.29)

$$\lambda_k(\mathbf{y}) = \tilde{\lambda}_k(\tilde{\mathbf{y}}) \quad \text{for } k \in \mathbb{N}, \quad (2.30)$$

and their eigenspaces coincide.

Moreover, for an eigenvalue $\tilde{\lambda}_k(\tilde{\mathbf{y}})$ (2.29), using (2.30) in the inequality (2.25) we have

$$|\tilde{\lambda}_k(\tilde{\mathbf{y}}) - \tilde{\lambda}_k(\tilde{\mathbf{y}}')| \leq C \max \left(\|a(\tilde{\mathbf{y}}/\boldsymbol{\alpha}) - a(\tilde{\mathbf{y}}'/\boldsymbol{\alpha})\|_{L^\infty(D)}, \|b(\tilde{\mathbf{y}}/\boldsymbol{\alpha}) - b(\tilde{\mathbf{y}}'/\boldsymbol{\alpha})\|_{L^\infty(D)} \right),$$

which after expanding the coefficients and using the triangle inequality becomes

$$|\tilde{\lambda}_k(\tilde{\mathbf{y}}) - \tilde{\lambda}_k(\tilde{\mathbf{y}}')| \leq \underbrace{C \left(\sum_{j=1}^{\infty} \frac{1}{\alpha_j} \max \left(\|a_j\|_{L^\infty(D)}, \|b_j\|_{L^\infty(D)} \right) \right)}_{\tilde{C}_{\text{Lip}}} \|\tilde{\mathbf{y}} - \tilde{\mathbf{y}}'\|_{\ell^\infty},$$

where \tilde{C}_{Lip} is clearly independent of $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{y}}'$. Now by (2.27) together with Assumption A1, we have

$$\sum_{j=1}^{\infty} \frac{\|a_j\|_{L^\infty(D)}}{\alpha_j} \leq \sum_{j=1}^{\infty} \|a_j\|_{L^\infty(D)}^{1-\varepsilon} = \sum_{j=1}^{\infty} \|a_j\|_{L^\infty(D)}^p < \infty,$$

with the analogous estimate for $\sum_{j=1}^{\infty} \|b_j\|_{L^\infty(D)}/\alpha_j$. Thus, $\tilde{C}_{\text{Lip}} < \infty$ and hence the reparametrised eigenvalues are continuous on $U(\boldsymbol{\alpha})$.

It immediately follows that the spectral gap $\tilde{\lambda}_2(\tilde{\mathbf{y}}) - \tilde{\lambda}_1(\tilde{\mathbf{y}})$ is also continuous on $U(\boldsymbol{\alpha})$, which by Lemma 2.2 is a compact subset of ℓ^∞ . Therefore, the non-zero minimum is attained giving that the spectral gap $\tilde{\lambda}_2(\tilde{\mathbf{y}}) - \tilde{\lambda}_1(\tilde{\mathbf{y}})$ is uniformly positive. Finally, because there is equality between the original and reparametrised eigenvalues (2.30) the result holds for the original problem over all $\mathbf{y} \in U$. \square

Remark 2.5 An explicit bound on the spectral gap can be obtained by assuming tighter restrictions on the coefficients. For example, if $a \equiv 1 \equiv c$ and b is weakly convex then [2] gives an explicit lower bound on the fundamental gap. Alternatively, using the upper and lower bounds on the eigenvalues (2.14), we can determine restrictions on a_{\min} and a_{\max} such that the gap is bounded away from 0. Explicitly,

$$\frac{a_{\min}}{a_{\max}} > \sqrt{\frac{\chi_1 + 1}{\chi_2}},$$

then $\lambda_2(\mathbf{y}) - \lambda_1(\mathbf{y}) \geq \underline{\lambda}_2 - \overline{\lambda}_1 > 0$.

2.3 Finite element discretisation

To approximate eigenpairs $(\lambda(\mathbf{y}), u(\mathbf{y}))$ we introduce a collection of finite element (FE) subspaces $V_h \subset V$ with dimension M_h , each of which is associated with a conforming triangulation \mathcal{T}_h of the domain D and a basis $(\phi_{h,i})_{i=1}^{M_h}$. The parameter $h = \max\{\text{diam}(\tau) : \tau \in \mathcal{T}_h\}$ is called the *meshwidth*. The method works for very general spaces V_h , however to fully exploit higher rates of convergence, stronger assumptions on the regularity of the coefficients and the domain would be required. As such, in the current paper we restrict our attention to piecewise linear finite elements, that is, each V_h is the space of continuous functions that are linear on the elements of \mathcal{T}_h and vanish on the boundary ∂D . It is well-known that, for $t \in (0, 1]$, the best approximation error for the space Z^t (as defined in (2.17) and (2.18)) by functions in V_h satisfies

$$\inf_{w_h \in V_h} \|v - w_h\|_V \leq Ch^t \|v\|_{Z^t} \quad \text{for all } v \in Z^t. \quad (2.31)$$

For each $\mathbf{y} \in U$ the discrete eigenvalue problem is to find $\lambda_h(\mathbf{y}) \in \mathbb{R}$ and $u_h(\mathbf{y}) \in V_h$ satisfying

$$\begin{aligned}\mathcal{A}(\mathbf{y}; u_h(\mathbf{y}), v) &= \lambda_h(\mathbf{y})\mathcal{M}(u_h(\mathbf{y}), v) \quad \text{for all } v \in V_h, \\ \|u_h(\mathbf{y})\|_{\mathcal{M}} &= 1.\end{aligned}\tag{2.32}$$

For each $\mathbf{y} \in U$, the discrete eigenvalue problem (2.32) admits M_h eigenvalues

$$0 < \lambda_{1,h}(\mathbf{y}) \leq \lambda_{2,h}(\mathbf{y}) \leq \dots \leq \lambda_{M_h,h}(\mathbf{y}),$$

with corresponding eigenfunctions

$$u_{1,h}(\mathbf{y}), u_{2,h}(\mathbf{y}), \dots, u_{M_h,h}(\mathbf{y}) \in V_h.$$

For each fixed k , the k th finite element eigenvalue $\lambda_{k,h}(\mathbf{y})$ converges from above to the k th eigenvalue of (2.7), i.e., for each k ,

$$\lambda_{k,h}(\mathbf{y}) \rightarrow \lambda_k(\mathbf{y}) \quad \text{as } h \rightarrow 0, \quad \text{with } \lambda_{k,h}(\mathbf{y}) \geq \lambda_k(\mathbf{y}) \quad \text{for all } h > 0,$$

and the corresponding FE eigenfunctions $(u_{k,h}(\mathbf{y}))_{k=1}^{M_h}$ satisfy

$$\text{dist}_V(u_{k,h}(\mathbf{y}), E(\mathbf{y}, \lambda_k(\mathbf{y}))) \rightarrow 0 \quad \text{as } h \rightarrow 0,$$

where $\text{dist}_V(v, \mathcal{P})$ is the distance of $v \in V$ from the subspace $\mathcal{P} \subset V$

$$\text{dist}_V(v, \mathcal{P}) := \inf_{w \in \mathcal{P}} \|v - w\|_V.$$

The classical results on FE error estimates for eigenproblems are found in [5–7]; however, we cannot simply use these results verbatim since their constants depend in complex and often hidden ways on the eigenvalues and eigenvalue gaps. For us, this means that they depend on \mathbf{y} , and so care must be taken to ensure that the constants do not become unbounded for some $\mathbf{y} \in U$. Theorem 2.6 below quantifies the FE convergence and in the proof we show that all constants are independent of \mathbf{y} . The proof is rather long and technical, and as such is left for the appendix.

Theorem 2.6 *Let $\mathbf{y} \in U$ and suppose that Assumption A1 holds. Then for $h > 0$ sufficiently small*

$$|\lambda_1(\mathbf{y}) - \lambda_{1,h}(\mathbf{y})| \leq C_1 h^2, \tag{2.33}$$

and $u_{1,h}(\mathbf{y}) \in E(\mathbf{y}, \lambda_{1,h}(\mathbf{y}))$ can be chosen such that

$$\|u_1(\mathbf{y}) - u_{1,h}(\mathbf{y})\|_V \leq C_2 h. \tag{2.34}$$

Moreover, for $\mathcal{G} \in H^{-1+t}(D)$ with $t \in [0, 1]$

$$|\mathcal{G}(u_1(\mathbf{y})) - \mathcal{G}(u_{1,h}(\mathbf{y}))| \leq C_3 h^{1+t}, \tag{2.35}$$

and $C_1, C_2, C_3 > 0$ are independent of \mathbf{y} .

2.4 Quasi-Monte Carlo methods

In this section, we give a brief overview of quasi-Monte Carlo (QMC) rules and the analysis of the resulting integration error. For more details on QMC point sets and theory, see [11].

QMC methods are equal-weight quadrature rules that can be used for the approximation of integrals over the (translated) s -dimensional unit cube

$$\mathcal{I}_s f := \int_{[-\frac{1}{2}, \frac{1}{2}]^s} f(\mathbf{y}) d\mathbf{y},$$

where the dimensionality s is high.

In this paper we use a class of QMC rules called randomly shifted rank-1 lattice rules, where the points are constructed using a *generating vector* $\mathbf{z} \in \mathbb{N}^s$ and a *shift* Δ , which is uniformly distributed on $[0, 1]^s$. Specifically, we have

$$\mathcal{Q}_{N,s} f := \frac{1}{N} \sum_{k=0}^{N-1} f \left(\left\{ \frac{k\mathbf{z}}{N} + \Delta \right\} - \frac{1}{2} \right),$$

where the braces denote taking the fractional part of each component and we have subtracted the vector $\frac{1}{2} := (\frac{1}{2}, \dots, \frac{1}{2})$ to map the points from $[0, 1]^s$ to $[-\frac{1}{2}, \frac{1}{2}]^s$. In practice, the advantages of random shifting are threefold: the final approximation is an unbiased estimate of the integral; using multiple shifts provides a practical estimate of root-mean-square (RMS) error; and the construction of a good lattice rule is simplified by the randomisation.

The error analysis of randomly shifted lattice rules requires the integrand belong to a weighted Sobolev space such as one of those first introduced in [35]. The s -dimensional weighted Sobolev space, denoted by $\mathcal{W}_{s,\gamma}$, is the space of functions with square-integrable mixed first derivatives and a norm which depends on a family of positive real numbers called weights. For each $\mathbf{u} \subseteq \{1, \dots, s\}$, the weight, denoted $\gamma_{\mathbf{u}}$, measures the “importance” of the subset of variables y_j with $j \in \mathbf{u}$. We let the entire collection of weights be denoted by γ .

In this paper we equip $\mathcal{W}_{s,\gamma}$ with the “unanchored” weighted norm. To define it we require the following notation: let $\{1 : s\} := \{1, \dots, s\}$, $\mathbf{y}_{\mathbf{u}} := (y_j)_{j \in \mathbf{u}}$, $\mathbf{y}_{-\mathbf{u}} := (y_j)_{j \in \{1:s\} \setminus \mathbf{u}}$ and let $\partial^{|\mathbf{u}|}/\partial \mathbf{y}_{\mathbf{u}} := \prod_{j \in \mathbf{u}} (\partial/\partial y_j)$ denote the first order mixed partial derivative with respect to the variables $\mathbf{y}_{\mathbf{u}}$. Now, let the norm (squared) of $f \in \mathcal{W}_{s,\gamma}$ be given by

$$\|f\|_{s,\gamma}^2 = \sum_{\mathbf{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathbf{u}}} \int_{[-\frac{1}{2}, \frac{1}{2}]^{|\mathbf{u}|}} \left(\int_{[-\frac{1}{2}, \frac{1}{2}]^{s-|\mathbf{u}|}} \frac{\partial^{|\mathbf{u}|}}{\partial \mathbf{y}_{\mathbf{u}}} f(\mathbf{y}) d\mathbf{y}_{-\mathbf{u}} \right)^2 d\mathbf{y}_{\mathbf{u}}. \quad (2.36)$$

Good generating vectors \mathbf{z} can be efficiently constructed using the *Fast CBC algorithm*, see [34] for CBC and [29,30] for “its acceleration” (or “fast CBC”). It has been shown, see e.g. [11, Theorem 5.10], that the RMS error of such a QMC approximation is bounded above by

$$\begin{aligned} & \sqrt{\mathbb{E}_{\Delta} \left(|\mathcal{I}_s f - Q_{N,s} f|^2 \right)} \\ & \leq \left(\frac{1}{\varphi(N)} \sum_{\emptyset \neq u \subseteq \{1:s\}} \gamma_u^\eta \left(\frac{2\zeta(2\eta)}{(2\pi^2)^\eta} \right)^{|u|} \right)^{\frac{1}{2\eta}} \|f\|_{s,\gamma} \quad \text{for all } \eta \in (\frac{1}{2}, 1], \end{aligned} \quad (2.37)$$

where the subscript Δ in \mathbb{E}_{Δ} indicates that the expectation is taken with respect to the (uniformly distributed) random shift, $\varphi(N) := |\{1 \leq \xi \leq N : \gcd(\xi, N) = 1\}|$ is the Euler totient function, and $\zeta(x) := \sum_{k=1}^{\infty} k^{-x}$ for $x > 1$ is the Riemann zeta function. In particular, if N is prime then $\varphi(N) = N - 1$.

3 Parametric regularity

In this section we examine the regularity with respect to y of the minimal eigenpair $(\lambda_1(y), u_1(y))$ of the variational eigenproblem (2.7). The results we obtain show that $\lambda_1(y)$ belongs to the weighted space $\mathcal{W}_{s,\gamma}$ with norm defined in (2.36). This is required for the analysis of the QMC error in approximating $\mathbb{E}_y[\lambda_1]$. Also, to obtain an *a priori* bound on the QMC error we require a bound on the norm of $\lambda_1(y)$ in $\mathcal{W}_{s,\gamma}$, hence we must bound its mixed first derivatives, see Lemma 3.4. There we use the bounds on the spectral gap obtained in § 2.2, and present results not only for the first-order mixed derivatives but also for higher-order derivatives.

We begin with the following coercive-type estimate, which is required in order to bound the norm of the derivatives of the eigenfunction.

Lemma 3.1 *Let Assumption A1 hold. Then for each $y \in U$ and $\lambda \in \mathbb{R}$, define $\mathcal{A}_\lambda^{\text{sh}}(y; \cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ to be the shifted bilinear form given by*

$$\mathcal{A}_\lambda^{\text{sh}}(y; u, v) := \mathcal{A}(y; u, v) - \lambda \mathcal{M}(u, v), \quad (3.1)$$

with \mathcal{A} and \mathcal{M} defined by (2.4) and (2.5), respectively. Restricted to the \mathcal{M} -orthogonal complement of the eigenspace corresponding to $\lambda_1(y)$, denoted by $E(y, \lambda_1(y))^\perp$, the $\lambda_1(y)$ -shifted bilinear form is uniformly coercive in y , i.e.,

$$\mathcal{A}_{\lambda_1(y)}^{\text{sh}}(y; v, v) \geq C_{\text{gap}} \|v\|_V^2 \quad \text{for all } v \in E(y, \lambda_1(y))^\perp, \quad (3.2)$$

where, for δ as in Proposition 2.4,

$$C_{\text{gap}} := \frac{a_{\min} \delta}{\bar{\lambda}_2}.$$

Proof Since the eigenfunctions $(u_k(y))_{k \in \mathbb{N}}$ form a basis in V that is orthonormal with respect to the inner product \mathcal{M} , for $v \in E(y, \lambda_1(y))^\perp$, letting $v_k(y) := \mathcal{M}(v, u_k(y))u_k(\cdot, y)$ for $k = 1, 2, \dots$, we can write

$$v = \sum_{k=2}^{\infty} v_k(y),$$

where we have used $v_1(y) = 0$ since $\mathcal{M}(v, u_1(y)) = 0$. Henceforth, we will suppress the dependence of the eigenvalues and v_k on y . For $v \in E(y, \lambda_1)^\perp$ we have

$$\mathcal{A}_{\lambda_1}^{\text{sh}}(y; v, v) = \mathcal{A}_{\lambda_1}^{\text{sh}} \left(y; \sum_{k=2}^{\infty} v_k, \sum_{\ell=2}^{\infty} v_{\ell} \right) = \sum_{k, \ell=2}^{\infty} (\mathcal{A}(y; v_k, v_{\ell}) - \lambda_1 \mathcal{M}(v_k, v_{\ell})).$$

Since all v_k are just scaled versions of u_k , they also satisfy the variational equation (2.3), so that $\mathcal{A}(y; v_k, v_{\ell}) = \lambda_k \mathcal{M}(v_k, v_{\ell})$ and they are orthogonal with respect to $\mathcal{M}(\cdot, \cdot)$, implying that $\mathcal{A}(y; v_k, v_{\ell}) = 0$ for $k \neq \ell$. Thus we can reduce the above double sum to

$$\begin{aligned} \mathcal{A}_{\lambda_1}^{\text{sh}}(y; v, v) &= \sum_{k=2}^{\infty} \left(\mathcal{A}(y; v_k, v_k) - \frac{\lambda_1}{\lambda_k} \mathcal{A}(y; v_k, v_k) \right) \\ &\geq \left(1 - \frac{\lambda_1}{\lambda_2} \right) \sum_{k=2}^{\infty} \mathcal{A}(y; v_k, v_k) = \left(1 - \frac{\lambda_1}{\lambda_2} \right) \sum_{k, \ell=2}^{\infty} \mathcal{A}(y; v_k, v_{\ell}) \\ &= \left(1 - \frac{\lambda_1}{\lambda_2} \right) \mathcal{A}(y; v, v) \geq a_{\min} \left(1 - \frac{\lambda_1}{\lambda_2} \right) \|v\|_V^2. \end{aligned}$$

The final bound, which is independent of y , follows from Proposition 2.4 and (2.14). \square

Remark 3.2 A similar estimate holds for the shifted bilinear form on $V_h \times V_h$, provided h is sufficiently small such that the FE eigenvalue gap is uniformly bounded from below. Indeed, we can write

$$\lambda_{2,h} - \lambda_{1,h} = (\lambda_{2,h} - \lambda_1) - (\lambda_{1,h} - \lambda_1),$$

and since the FE eigenvalues converge from above we can bound this from below by

$$\lambda_{2,h} - \lambda_{1,h} \geq \lambda_2 - \lambda_1 - |\lambda_1 - \lambda_{1,h}| \geq \delta - Ch^2.$$

with $C > 0$. The second inequality follows from Proposition 2.4 and Theorem 2.6. Thus, choosing h such that $Ch^2 < \delta$, or equivalently, taking $h < h_0$ with

$$h_0 := \left(\frac{\delta}{C} \right)^{\frac{1}{2}}, \quad (3.3)$$

is a sufficient condition for $\lambda_{2,h} - \lambda_{1,h} > 0$, and then Lemma 3.1 can be rewritten for the FE eigenproblem.

We also require the following technical lemma to handle some combinatorial factors that arise when bounding the derivatives.

Lemma 3.3 *Let $\epsilon \in (0, 1)$. For all $n \in \mathbb{N}$, the following bound holds:*

$$S_n(\epsilon) := \sum_{k=1}^{n-1} \binom{n}{k}^{-\epsilon} \leq C_\epsilon := \frac{2^{1-\epsilon}}{1-2^{-\epsilon}} \left(\frac{e^2}{\sqrt{2\pi}} \right)^\epsilon.$$

Proof By the symmetry of the binomial coefficient the sum can be bounded by

$$S_n(\epsilon) \leq 2 \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{k}^{-\epsilon} \leq 2 \left(\frac{e^2}{\sqrt{2\pi}} \right)^\epsilon \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \left(\frac{k^{k+\frac{1}{2}} (n-k)^{n-k+\frac{1}{2}}}{n^{n+\frac{1}{2}}} \right)^\epsilon,$$

where we used the following bounds given by Stirling's formula

$$\sqrt{2\pi n^{n+\frac{1}{2}}} e^{-n} \leq n! \leq e n^{n+\frac{1}{2}} e^{-n}.$$

Since $k \leq \frac{n}{2}$ we have the bound $k^{k+\frac{1}{2}} \leq (\frac{n}{2})^k \sqrt{k}$, which gives

$$\begin{aligned} S_n(\epsilon) &\leq 2 \left(\frac{e^2}{\sqrt{2\pi}} \right)^\epsilon \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \left(\frac{(\frac{n}{2})^k \sqrt{k} (n-k)^{n-k+\frac{1}{2}}}{n^{n+\frac{1}{2}}} \right)^\epsilon \\ &= 2 \left(\frac{e^2}{\sqrt{2\pi}} \right)^\epsilon \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \frac{1}{2^{\epsilon k}} \left(\sqrt{k} \left(1 - \frac{k}{n} \right)^{n-k+\frac{1}{2}} \right)^\epsilon. \end{aligned}$$

The next step is to show that for each term in the sum the factor occurring inside the brackets is always bounded above by 1. To proceed, for $n = 1, 2, 3, \dots$ we define the functions $R_n : [0, 1/2] \rightarrow \mathbb{R}$ by

$$R_n(x) := \sqrt{nx} (1-x)^{n(1-x)+\frac{1}{2}}.$$

We prove by induction that for $n = 1, 2, 3, \dots$

$$R_n(x) \leq 1 \quad \text{for all } x \in [0, \frac{1}{2}]. \tag{3.4}$$

For $n = 1$

$$R_1(x) = \sqrt{x} (1-x)^{1-x+\frac{1}{2}} \leq \sqrt{\frac{1}{2}} (1-x)^{1-x+\frac{1}{2}} \leq 1.$$

For $n \geq 1$ suppose $R_n(x) \leq 1$ and consider R_{n+1} . For x in the interval $[1/(n+1), 1/2]$

$$\begin{aligned} R_{n+1}(x) &= \sqrt{(n+1)x}(1-x)^{(n+1)(1-x)+\frac{1}{2}} = \sqrt{\frac{n+1}{n}}\sqrt{nx}(1-x)^{1-x}(1-x)^{n(1-x)+\frac{1}{2}} \\ &= (1-x)^{1-x}\sqrt{\frac{n+1}{n}}R_n(x) \leq (1-x)^{1-x}\sqrt{\frac{n+1}{n}}. \end{aligned}$$

To bound this from above we bound one x below by $1/(n+1)$ to give

$$R_{n+1}(x) \leq \left(1 - \frac{1}{n+1}\right)^{1-x} \sqrt{\frac{n+1}{n}} = \left(\frac{n}{n+1}\right)^{\frac{1}{2}-x} \leq 1.$$

And for $x \in [0, 1/(n+1)]$

$$R_{n+1}(x) \leq \sqrt{(n+1)\frac{1}{n+1}}(1-x)^{(n+1)(1-x)+\frac{1}{2}} = (1-x)^{(n+1)(1-x)+\frac{1}{2}} \leq 1.$$

Thus, for all $n = 1, 2, 3, \dots$ and $x \in [0, 1/2]$ we have $R_n(x) \leq 1$.

Returning to the sum S_n , since

$$\frac{k}{n} \in [0, \frac{1}{2}] \quad \text{for all } k \leq \frac{n}{2}, \quad \text{and} \quad R_n\left(\frac{k}{n}\right) = \sqrt{k} \left(1 - \frac{k}{n}\right)^{n-k+\frac{1}{2}}$$

we have, by (3.4),

$$\begin{aligned} S_n(\epsilon) &\leq 2 \left(\frac{e^2}{\sqrt{2\pi}}\right)^\epsilon \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \frac{R_n\left(\frac{k}{n}\right)^\epsilon}{(2^\epsilon)^k} \leq 2 \left(\frac{e^2}{\sqrt{2\pi}}\right)^\epsilon \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \frac{1}{(2^\epsilon)^k} \\ &\leq 2 \left(\frac{e^2}{\sqrt{2\pi}}\right)^\epsilon \sum_{k=1}^{\infty} \frac{1}{(2^\epsilon)^k} = \frac{2^{1-\epsilon}}{1-2^{-\epsilon}} \left(\frac{e^2}{\sqrt{2\pi}}\right)^\epsilon =: C_\epsilon, \end{aligned}$$

where we used the formula for the sum of a geometric series. \square

Lemma 3.4 below gives the bounds on the derivatives of λ_1 and u_1 required for our QMC error analysis. We prove the bounds for higher order mixed derivatives, which will be written in multi-index notation. Let $\nu = (\nu_j)_{j \in \mathbb{N}}$, with $\nu_j \in \mathbb{N} \cup \{0\}$, be a multi-index with only finitely many non-zero entries and define $|\nu| := \sum_{j \geq 1} \nu_j$. We call such a multi-index *admissible*, and let \mathcal{F} denote the set of all admissible multi-indexes. We will use ∂_y^ν to denote the mixed partial derivative where the element ν_j is the order of the derivative with respect to y_j . Operations between multi-indices are handled component wise. Thus, for $\mathbf{m} = (m_j)_{j \in \mathbb{N}}$, $\nu = (\nu_j)_{j \in \mathbb{N}}$ we use the following notation: $\nu! = \prod_{j \geq 1} \nu_j!$; $\nu - \mathbf{m} := (\nu_j - m_j)_{j \in \mathbb{N}}$; $\mathbf{m} \leq \nu$ if $m_j \leq \nu_j$ for all $j \in \mathbb{N}$; $\mathbf{m} < \nu$ if $\mathbf{m} \leq \nu$ and $\mathbf{m} \neq \nu$; $\binom{\nu}{\mathbf{m}} := \prod_{j \in \mathbb{N}} \binom{\nu_j}{m_j}$; and for $\alpha = (\alpha_j)_{j \in \mathbb{N}} \in \ell^\infty$ let $\alpha^\nu := \prod_{j \in \mathbb{N}} \alpha_j^{\nu_j}$. For $j \in \mathbb{N}$, the j th unit multi-index is denoted by e_j , that is, e_j is 1 in the j th position and 0 everywhere else.

Since the coefficients $a(\mathbf{x}, \mathbf{y})$ and $b(\mathbf{x}, \mathbf{y})$ in (1.3) are linear in the parameter \mathbf{y} , their derivatives are (suppressing the \mathbf{x}, \mathbf{y} dependence below)

$$\partial_y^v a = \begin{cases} a & \text{if } v = \mathbf{0}, \\ a_j & \text{if } v = e_j, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad \partial_y^v b = \begin{cases} b & \text{if } v = \mathbf{0}, \\ b_j & \text{if } v = e_j, \\ 0 & \text{otherwise.} \end{cases} \quad (3.5)$$

Lemma 3.4 *Let $\epsilon \in (0, 1)$, $v \in \mathcal{F}$ be a multi-index, and suppose that Assumption A1 holds. Then for all $\mathbf{y} \in U$ the corresponding derivative of the smallest eigenvalue of (2.7) is bounded by*

$$|\partial_y^v \lambda_1(\mathbf{y})| \leq \overline{\lambda_1} (|v|!)^{1+\epsilon} \beta^v, \quad (3.6)$$

and the norm of the derivative of the corresponding eigenfunction is similarly bounded by

$$\|\partial_y^v u_1(\mathbf{y})\|_V \leq \overline{u_1} (|v|!)^{1+\epsilon} \beta^v, \quad (3.7)$$

where $\overline{\lambda_1}$ and $\overline{u_1}$ are defined in (2.14) and (2.16), respectively. The sequence $\beta = (\beta_j)_{j \in \mathbb{N}}$ is defined by

$$\beta_j := C_\beta \max (\|a_j\|_{L^\infty(D)}, \|b_j\|_{L^\infty(D)}), \quad (3.8)$$

with $C_\beta > 0$ independent of \mathbf{y} , given by

$$C_\beta := \frac{1}{C_{\text{gap}}} \frac{a_{\min}^2 \overline{\lambda_1}}{a_{\max}^2 \underline{\lambda_1}} \left(\frac{3\overline{\lambda_1}}{2\underline{\lambda_1}} C_\epsilon + 1 \right), \quad (3.9)$$

where C_{gap} is as in Lemma 3.1, and C_ϵ is as in Lemma 3.3.

Proof For $v = \mathbf{0}$ the bounds (3.6) and (3.7) clearly hold by (2.14) and (2.16), respectively.

For $v \neq \mathbf{0}$, we will prove the two bounds by induction on $|v|$. To this end, we first obtain recursive bounds by differentiating the variational form (2.3) with respect to the stochastic parameters $\mathbf{y} \in U$, see (3.12) and (3.19), which will then be used to prove (3.6) and (3.7) inductively. From [1], we know that simple eigenpairs of (2.7) are analytic in \mathbf{y} , so the partial derivatives $\partial_y^v \lambda_1, \partial_y^v u_1$ exist and we further have $\partial_y^v u_1 \in V$. Hence, we can differentiate (2.3) with $\lambda = \lambda_1$ and $u = u_1$ using the Leibniz general product rule to obtain the following formula, which is true for all $v \in V$,

$$\begin{aligned} & \sum_{\mathbf{m} \leq \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} \left(- \left(\partial_y^{\mathbf{m}} \lambda_1(\mathbf{y}) \right) \int_D c(\mathbf{x}) \left(\partial_y^{\mathbf{v}-\mathbf{m}} u_1(\mathbf{x}, \mathbf{y}) \right) v(\mathbf{x}) \, d\mathbf{x} \right. \\ & + \int_D \left(\partial_y^{\mathbf{m}} a(\mathbf{x}, \mathbf{y}) \right) \nabla \left(\partial_y^{\mathbf{v}-\mathbf{m}} u_1(\mathbf{x}, \mathbf{y}) \right) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x} \\ & \left. + \int_D \left(\partial_y^{\mathbf{m}} b(\mathbf{x}, \mathbf{y}) \right) \left(\partial_y^{\mathbf{v}-\mathbf{m}} u_1 \right) v(\mathbf{x}) \, d\mathbf{x} \right) = 0. \end{aligned} \quad (3.10)$$

Henceforth, we consider $\mathbf{y} \in U$ to be fixed and will suppress the dependence of $a(\mathbf{x}, \mathbf{y}), b(\mathbf{x}, \mathbf{y}), c(\mathbf{x}), \lambda_1(\mathbf{y}), u_1(\mathbf{x}, \mathbf{y})$, and their respective derivatives, on \mathbf{x} and \mathbf{y} .

To obtain a bound on the derivatives of the eigenvalue, we take $v = u_1$ in (3.10). In this case, the $\mathbf{m} = \mathbf{0}$ term vanishes since (2.3) is satisfied for $\lambda = \lambda_1$ and $u = u_1$ with $v = u_1 \in V$ as a test function also. Separating out the $\partial_y^{\mathbf{v}} \lambda_1$ term and using $\|u_1\|_{\mathcal{M}} = 1$ gives

$$\begin{aligned} \partial_y^{\mathbf{v}} \lambda_1 &= - \sum_{\mathbf{0} \neq \mathbf{m} < \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} (\partial_y^{\mathbf{m}} \lambda_1) \int_D c \left(\partial_y^{\mathbf{v}-\mathbf{m}} u_1 \right) u_1 \\ &+ \sum_{j=1}^{\infty} v_j \left(\int_D a_j \nabla \left(\partial_y^{\mathbf{v}-\mathbf{e}_j} u_1 \right) \cdot \nabla u_1 + \int_D b_j \left(\partial_y^{\mathbf{v}-\mathbf{e}_j} u_1 \right) u_1 \right), \end{aligned} \quad (3.11)$$

since the terms involving higher-order derivatives of the coefficients are 0 (see (3.5)). Taking the absolute value then applying the triangle and Cauchy–Schwarz inequalities gives the upper bound

$$\begin{aligned} |\partial_y^{\mathbf{v}} \lambda_1| &\leq \sum_{\mathbf{0} \neq \mathbf{m} < \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} |\partial_y^{\mathbf{m}} \lambda_1| \|\partial_y^{\mathbf{v}-\mathbf{m}} u_1\|_{\mathcal{M}} \|u_1\|_{\mathcal{M}} \\ &+ \sum_{j=1}^{\infty} v_j \left(\|a_j\|_{L^\infty(D)} \|\partial_y^{\mathbf{v}-\mathbf{e}_j} u_1\|_V \|u_1\|_V + \|b_j\|_{L^\infty(D)} \|\partial_y^{\mathbf{v}-\mathbf{e}_j} u_1\|_{L^2(D)} \|u_1\|_{L^2(D)} \right). \end{aligned}$$

Then, by the equivalence of the norms $\|\cdot\|_{\mathcal{M}}$ and $\|\cdot\|_{L^2(D)}$ in (2.6), the Poincaré inequality (2.10), the upper bound (2.16) on $\|u_1\|_V$, and the normalisation of u_1 , we have

$$\begin{aligned} |\partial_y^{\mathbf{v}} \lambda_1| &\leq \sqrt{\frac{a_{\max}}{\chi_1}} \sum_{\mathbf{0} \neq \mathbf{m} < \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} |\partial_y^{\mathbf{m}} \lambda_1| \|\partial_y^{\mathbf{v}-\mathbf{m}} u_1\|_V \\ &+ \overline{u_1} \sum_{j=1}^{\infty} v_j \left(\|a_j\|_{L^\infty(D)} + \frac{1}{\chi_1} \|b_j\|_{L^\infty(D)} \right) \|\partial_y^{\mathbf{v}-\mathbf{e}_j} u_1\|_V. \end{aligned}$$

Defining β_j as in (3.8) but leaving $C_\beta > 0$ to be specified later, we obtain

$$|\partial_y^v \lambda_1| \leq \sqrt{\frac{a_{\max}}{\chi_1}} \sum_{\mathbf{0} \neq \mathbf{m} < v} \binom{\mathbf{v}}{\mathbf{m}} |\partial_y^{\mathbf{m}} \lambda_1| \|\partial_y^{\mathbf{v}-\mathbf{m}} u_1\|_V + \overline{u_1} \left(1 + \frac{1}{\chi_1}\right) \sum_{j=1}^{\infty} v_j \frac{\beta_j}{C_{\beta}} \|\partial_y^{\mathbf{v}-e_j} u_1\|_V, \quad (3.12)$$

which depends only on the lower order derivatives of both λ_1 and u_1 .

Substituting $v = \partial_y^v u_1$ into (3.10) to obtain a similar bound on the derivatives of the eigenfunction will not work, because in (3.10) the bilinear form acting on $\partial_y^v u_1$ (exactly $\mathcal{A}_{\lambda_1}^{\text{sh}}$ from Lemma 3.1) is not coercive on the whole domain $V \times V$. The way around this is to expand $\partial_y^v u_1$ in the eigenbasis and then utilise the estimate in Lemma 3.1. We write $\partial_y^v u_1$ as

$$\partial_y^v u_1 = \sum_{k \in \mathbb{N}} \mathcal{M}(\partial_y^v u_1, u_k) u_k = \mathcal{M}(\partial_y^v u_1, u_1) u_1 + \tilde{v}, \quad (3.13)$$

so that $\tilde{v} \in E(\mathbf{y}, \lambda_1(\mathbf{y}))^\perp$ is the \mathcal{M} -orthogonal projection of $\partial_y^v u_1$ onto $E(\mathbf{y}, \lambda_1(\mathbf{y}))^\perp$. Applying the triangle inequality to this decomposition and then using (2.16) we can bound the norm by

$$\|\partial_y^v u_1\|_V \leq \overline{u_1} |\mathcal{M}(\partial_y^v u_1, u_1)| + \|\tilde{v}\|_V. \quad (3.14)$$

Hence, it remains to bound $|\mathcal{M}(\partial_y^v u_1, u_1)|$ and $\|\tilde{v}\|_V$. For the former, since $\mathcal{M}(u_1, u_1) = 1$ we have

$$0 = \partial_y^v \mathcal{M}(u_1, u_1) = \sum_{\mathbf{m} \leq \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} \mathcal{M}(\partial_y^{\mathbf{m}} u_1, \partial_y^{\mathbf{v}-\mathbf{m}} u_1).$$

By separating out the $\mathbf{m} = \mathbf{0}$ and $\mathbf{m} = \mathbf{v}$ terms, which are equal by symmetry, we obtain

$$|\mathcal{M}(\partial_y^v u_1, u_1)| = \left| -\frac{1}{2} \sum_{\mathbf{0} \neq \mathbf{m} < \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} \mathcal{M}(\partial_y^{\mathbf{m}} u_1, \partial_y^{\mathbf{v}-\mathbf{m}} u_1) \right| \quad (3.15)$$

$$\leq \frac{a_{\max}}{2\chi_1} \sum_{\mathbf{0} \neq \mathbf{m} < \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} \|\partial_y^{\mathbf{m}} u_1\|_V \|\partial_y^{\mathbf{v}-\mathbf{m}} u_1\|_V, \quad (3.16)$$

where we used the Cauchy–Schwarz inequality, the norm equivalence (2.6) and then the Poincaré inequality (2.10) to obtain V -norms.

For the V -norm of \tilde{v} , we let $v = \tilde{v}$ in (3.10) and separate out the $\mathbf{m} = \mathbf{0}$ term to give

$$\begin{aligned} \mathcal{A}(\partial_y^v u_1, \tilde{v}) - \lambda_1 \mathcal{M}(\partial_y^v u_1, \tilde{v}) &= \sum_{\mathbf{0} \neq \mathbf{m} < \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} (\partial_y^{\mathbf{m}} \lambda_1) \int_D c(\partial_y^{\mathbf{v}-\mathbf{m}} u_1) \tilde{v} \\ &\quad - \sum_{j=1}^{\infty} v_j \left(\int_D a_j \nabla(\partial_y^{\mathbf{v}-e_j} u_1) \cdot \nabla \tilde{v} + \int_D b_j(\partial_y^{\mathbf{v}-e_j} u_1) \tilde{v} \right), \end{aligned} \quad (3.17)$$

where the $\mathbf{m} = \mathbf{v}$ term vanishes from the right-hand side since \tilde{v} is orthogonal to u_1 . Decomposing $\partial_y^{\mathbf{v}} u_1$ as in (3.13) and then expanding, the left-hand side of (3.17) becomes

$$\begin{aligned}\text{LHS of (3.17)} &= \mathcal{M}(\partial_y^{\mathbf{v}} u_1, u_1)(\mathcal{A}(u_1, \tilde{v}) - \lambda_1 \mathcal{M}(u_1, \tilde{v})) + \mathcal{A}(\tilde{v}, \tilde{v}) - \lambda_1 \mathcal{M}(\tilde{v}, \tilde{v}) \\ &= \mathcal{A}(\tilde{v}, \tilde{v}) - \lambda_1 \mathcal{M}(\tilde{v}, \tilde{v}) \geq C_{\text{gap}} \|\tilde{v}\|_V^2,\end{aligned}$$

where the first term on the first line is 0 by (2.3) with \tilde{v} as a test function. The lower bound follows by the coercivity estimate (3.2) in Lemma 3.1, since $\tilde{v} \in E(y, \lambda_1(y))^\perp$.

The right-hand side of (3.17) can be bounded from above as for (3.11) to obtain

$$\begin{aligned}C_{\text{gap}} \|\tilde{v}\|_V^2 &\leq \frac{a_{\max}}{\chi_1} \sum_{\mathbf{0} \neq \mathbf{m} < \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} |\partial_y^{\mathbf{m}} \lambda_1| \|\partial_y^{\mathbf{v}-\mathbf{m}} u_1\|_V \|\tilde{v}\|_V \\ &\quad + \sum_{j=1}^{\infty} v_j \left(\|a_j\|_{L^\infty(D)} + \frac{1}{\chi_1} \|b_j\|_{L^\infty(D)} \right) \|\partial_y^{\mathbf{v}-\mathbf{e}_j} u_1\|_V \|\tilde{v}\|_V.\end{aligned}$$

Dividing through by $C_{\text{gap}} \|\tilde{v}\|_V$ and using the definition of β_j in (3.8), again leaving $C_{\boldsymbol{\beta}} > 0$ to be specified later, we obtain

$$\begin{aligned}\|\tilde{v}\|_V &\leq \frac{1}{C_{\text{gap}}} \left(\frac{a_{\max}}{\chi_1} \sum_{\mathbf{0} \neq \mathbf{m} < \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} |\partial_y^{\mathbf{m}} \lambda_1| \|\partial_y^{\mathbf{v}-\mathbf{m}} u_1\|_V \right. \\ &\quad \left. + \left(1 + \frac{1}{\chi_1} \right) \sum_{j=1}^{\infty} v_j \frac{\beta_j}{C_{\boldsymbol{\beta}}} \|\partial_y^{\mathbf{v}-\mathbf{e}_j} u_1\|_V \right). \quad (3.18)\end{aligned}$$

Substituting the two bounds (3.15) and (3.18) into (3.14), the norm of the derivative of the eigenfunction is bounded above by

$$\begin{aligned}\|\partial_y^{\mathbf{v}} u_1\|_V &\leq \overline{u_1} \frac{a_{\max}}{2\chi_1} \sum_{\mathbf{0} \neq \mathbf{m} < \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} \|\partial_y^{\mathbf{m}} u_1\|_V \|\partial_y^{\mathbf{v}-\mathbf{m}} u_1\|_V \\ &\quad + \frac{a_{\max}}{\chi_1 C_{\text{gap}}} \sum_{\mathbf{0} \neq \mathbf{m} < \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} |\partial_y^{\mathbf{m}} \lambda_1| \|\partial_y^{\mathbf{v}-\mathbf{m}} u_1\|_V \\ &\quad + \frac{1}{C_{\text{gap}}} \left(1 + \frac{1}{\chi_1} \right) \sum_{j=1}^{\infty} v_j \frac{\beta_j}{C_{\boldsymbol{\beta}}} \|\partial_y^{\mathbf{v}-\mathbf{e}_j} u_1\|_V. \quad (3.19)\end{aligned}$$

We are now ready to prove the bounds (3.6) and (3.7) by induction. To avoid any blow-up in the inductive step we require tighter constants than $\overline{\lambda_1}$ and $\overline{u_1}$. Thus we proceed to prove that, for $\mathbf{v} \neq \mathbf{0}$,

$$|\partial_y^v \lambda_1(y)| \leq C_1 (|v|!)^{1+\epsilon} \beta^v, \quad \text{and} \quad (3.20)$$

$$\|\partial_y^v u_1(y)\|_V \leq C_2 (|v|!)^{1+\epsilon} \beta^v, \quad (3.21)$$

where

$$C_1 := \frac{\overline{\lambda_1}}{a_{\min}} \left(1 + \frac{1}{\chi_1}\right) \frac{1}{C_\beta}, \quad \text{and}$$

$$C_2 := \frac{\overline{u_1}}{C_{\text{gap}}} \left(1 + \frac{1}{\chi_1}\right) \frac{1}{C_\beta},$$

with $C_\beta > 0$ still to be specified later to ensure that $C_1 \leq \overline{\lambda_1}$ and $C_2 \leq \overline{u_1}$.

Since the bounds (3.12) and (3.19) are true for all $v \neq \mathbf{0}$, and these bounds do not involve the $v = \mathbf{0}$ cases, we will use them to establish the base step of the induction $|v| = 1$. Letting $v = e_i$ in (3.12) and (3.19), and then using the bounds in (2.14) and (2.16) gives

$$\begin{aligned} |\partial_y^{e_i} \lambda_1| &\leq \overline{u_1}^2 \left(1 + \frac{1}{\chi_1}\right) \frac{\beta_i}{C_\beta} = \frac{\overline{\lambda_1}}{a_{\min}} \left(1 + \frac{1}{\chi_1}\right) \frac{\beta_i}{C_\beta}, \quad \text{and} \\ \|\partial_y^{e_i} u_1\|_V &\leq \frac{\overline{u_1}}{C_{\text{gap}}} \left(1 + \frac{1}{\chi_1}\right) \frac{\beta_i}{C_\beta}, \end{aligned}$$

as required.

For the inductive step for the eigenvalue derivative bound, suppose that $|v| \geq 2$ and that the bounds (3.20) and (3.21) hold for all multi-indices of order $< |v|$. Substituting the induction assumptions (3.20) and (3.21) into (3.12) and then factoring out $C_1 \beta^v$ gives

$$\begin{aligned} |\partial_y^v \lambda_1| &\leq \sqrt{\frac{a_{\max}}{\chi_1}} \sum_{\mathbf{0} \neq \mathbf{m} < v} \binom{v}{\mathbf{m}} C_1 (|\mathbf{m}|!)^{1+\epsilon} \beta^{\mathbf{m}} \cdot C_2 (|v - \mathbf{m}|!)^{1+\epsilon} \beta^{v-\mathbf{m}} \\ &\quad + \overline{u_1} \left(1 + \frac{1}{\chi_1}\right) \sum_{j \in \mathbb{N}} v_j \frac{\beta_j}{C_\beta} \cdot C_2 [(|v| - 1)!]^{1+\epsilon} \beta^{v-e_j} \\ &\leq C_1 \beta^v \left(\sqrt{\frac{a_{\max}}{\chi_1}} C_2 \sum_{\mathbf{0} \neq \mathbf{m} < v} \binom{v}{\mathbf{m}} (|\mathbf{m}|!)^{1+\epsilon} (|v - \mathbf{m}|!)^{1+\epsilon} \right. \\ &\quad \left. + \overline{u_1} \left(1 + \frac{1}{\chi_1}\right) \frac{C_2}{C_1 C_\beta} |v| [(|v| - 1)!]^{1+\epsilon} \right). \end{aligned}$$

Using the identity $\sum_{\mathbf{m} \leq v, |\mathbf{m}|=k} \binom{v}{\mathbf{m}} = \binom{|v|}{k}$ along with Lemma 3.3, we can bound the sum as follows

$$\begin{aligned}
& \sum_{\mathbf{0} \neq \mathbf{m} < \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} (|\mathbf{m}|!)^{1+\epsilon} (|\mathbf{v} - \mathbf{m}|!)^{1+\epsilon} \\
&= \sum_{k=1}^{|\mathbf{v}|-1} \left[\sum_{\mathbf{m} \leq \mathbf{v}, |\mathbf{m}|=k} \binom{\mathbf{v}}{\mathbf{m}} \right] (k!)^{1+\epsilon} [(|\mathbf{v}| - k)!]^{1+\epsilon} \\
&= (|\mathbf{v}|!)^{1+\epsilon} \sum_{k=1}^{|\mathbf{v}|-1} \binom{|\mathbf{v}|}{k}^{-\epsilon} \leq C_\epsilon (|\mathbf{v}|!)^{1+\epsilon}. \tag{3.22}
\end{aligned}$$

Substituting this into the bound on $|\partial_y^{\mathbf{v}} \lambda_1|$ yields

$$|\partial_y^{\mathbf{v}} \lambda_1| \leq C_1 (|\mathbf{v}|!)^{1+\epsilon} \beta^{\mathbf{v}} \left[\sqrt{\frac{a_{\max}}{\chi_1}} C_2 C_\epsilon + \overline{u_1} \left(1 + \frac{1}{\chi_1} \right) \frac{C_2}{C_1 C_\beta} \right].$$

Substituting in the values for C_1 and C_2 , and then using (2.14) and (2.16) the expression in between the square brackets simplifies to

$$\frac{1}{C_\beta C_{\text{gap}}} \frac{a_{\min}^2 \overline{\lambda_1}}{a_{\max}^2 \underline{\lambda_1}} \left(C_\epsilon \sqrt{\frac{\overline{\lambda_1}}{\underline{\lambda_1}}} + 1 \right) \tag{3.23}$$

and we will later specify C_β to ensure that this expression is bounded by 1, thus giving the required result (3.20).

For the eigenfunction derivative bounds, substituting the induction hypotheses (3.20) and (3.21) into (3.19)

$$\begin{aligned}
\|\partial_y^{\mathbf{v}} u_1\|_V &\leq \frac{a_{\max}}{2\chi_1} \overline{u_1} \sum_{\mathbf{0} \neq \mathbf{m} < \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} C_2 (|\mathbf{m}|!)^{1+\epsilon} \beta^{\mathbf{m}} \cdot C_2 (|\mathbf{v} - \mathbf{m}|!)^{1+\epsilon} \beta^{\mathbf{v}-\mathbf{m}} \\
&\quad + \frac{a_{\max}}{\chi_1 C_{\text{gap}}} \sum_{\mathbf{0} \neq \mathbf{m} < \mathbf{v}} \binom{\mathbf{v}}{\mathbf{m}} C_1 (|\mathbf{m}|!)^{1+\epsilon} \beta^{\mathbf{m}} \cdot C_2 (|\mathbf{v} - \mathbf{m}|!)^{1+\epsilon} \beta^{\mathbf{v}-\mathbf{m}} \\
&\quad + \frac{1}{C_{\text{gap}}} \left(1 + \frac{1}{\chi_1} \right) \sum_{j=1}^{\infty} v_j \frac{\beta_j}{C_\beta} C_2 [(\mathbf{v} - \mathbf{e}_j)!]^{1+\epsilon} \beta^{\mathbf{v}-\mathbf{e}_j}.
\end{aligned}$$

Factoring out $C_2 \beta^{\mathbf{v}}$ and using (3.22) this becomes

$$\|\partial_y^{\mathbf{v}} u_1\|_V \leq C_2 (|\mathbf{v}|!)^{1+\epsilon} \beta^{\mathbf{v}} \left[\frac{a_{\max}}{2\chi_1} \overline{u_1} C_2 C_\epsilon + \frac{a_{\max}}{\chi_1 C_{\text{gap}}} C_1 C_\epsilon + \frac{1}{C_{\text{gap}}} \left(1 + \frac{1}{\chi_1} \right) \frac{1}{C_\beta} \right],$$

Substituting in C_1 and C_2 , and then using again (2.14) and (2.16) the expression in between the square brackets simplifies to

$$\frac{1}{C_\beta C_{\text{gap}}} \frac{a_{\min}^2 \bar{\lambda}_1}{a_{\max}^2 \bar{\lambda}_1} \left(\frac{3\bar{\lambda}_1}{2\bar{\lambda}_1} C_\epsilon + 1 \right). \quad (3.24)$$

We now define C_β as in (3.9), so that the expression in (3.24) is exactly 1, thus proving the required bound for the eigenfunction (3.21), and ensuring also that the expression in (3.23) is bounded by 1 as required. This completes the induction proof for (3.20) and (3.21) for all $\mathbf{v} \neq \mathbf{0}$.

With this definition of C_β it can be verified that $C_1 \leq \bar{\lambda}_1$ and $C_2 \leq \bar{u}_1$ as required. Hence we have also proved (3.6) and (3.7) for all $\mathbf{v} \in \mathcal{F}$. \square

Remark 3.5 Since $V_h \subset V$, for h sufficiently small similar results can be proved analogously for the FE approximations, with the constants replaced by their FE counterparts $C_{\beta,h}, \bar{\lambda}_{1,h}, \bar{u}_{1,h}$.

4 Error analysis

Since we are only interested in the fundamental eigenpair, to aid in the notation we drop the subscript 1 and define $(\lambda, u) := (\lambda_1, u_1)$. Henceforth, with a slight abuse of notation, we will use combinations of the subscripts s, h, N to denote, respectively, truncating the stochastic dimension to s variables, a FE approximation with meshwidth h and a lattice rule approximation with N points. Also, for the dimension truncation we will denote the truncated parameter vector by $\mathbf{y}_s := (y_1, y_2, \dots, y_s)$.

Clearly, if the functions $(a_j)_{j \geq 1}, (b_j)_{j \geq 1}$ satisfy Assumption A1.3 then the sequence β defined by (3.8) is summable with the same p . Also, we will henceforth assume that β is ordered such that $\beta_1 \geq \beta_2 \geq \dots$.

4.1 Dimension truncation error

Here we present bounds on the error of the truncated eigenvalue, $\lambda_s(\mathbf{y}_s) := \lambda(\mathbf{y}_s; \mathbf{0})$, and the truncated eigenfunction, $u_s(\mathbf{y}) := u_s(\mathbf{y}_s; \mathbf{0})$, for both a given \mathbf{y} (the strong error) and the expected value (the weak error). To prove these estimates we will make extensive use of Taylor series expansions in the variables $(y_j)_{j > s}$ about $\mathbf{0}$ with integral remainders (as in [10]). However, motivated by the splitting strategy in [17], we use a higher order Taylor series expansion to obtain the same rate as in [17], which is an extra order of convergence for the weak error when compared with [10, 26].

Theorem 4.1 Suppose that Assumption A1 holds with $p \in (0, 1)$. There exists constants $C_1, C_2, C_3, C_4 > 0$ such that if $s \in \mathbb{N}$ is sufficiently large, then for all $\mathbf{y} \in U$ the strong truncation error of the minimal eigenpair is bounded by

$$|\lambda(\mathbf{y}) - \lambda_s(\mathbf{y}_s)| \leq C_1 s^{-1/p+1}, \quad (4.1)$$

$$\|u(\mathbf{y}) - u_s(\mathbf{y}_s)\|_V \leq C_2 s^{-1/p+1}. \quad (4.2)$$

The weak truncation error is bounded by

$$|\mathbb{E}_y [\lambda - \lambda_s]| \leq C_3 s^{-2/p+1}, \quad (4.3)$$

and for $\mathcal{G} \in V^*$

$$|\mathbb{E}_y [\mathcal{G}(u) - \mathcal{G}(u_s)]| \leq C_4 s^{-2/p+1}. \quad (4.4)$$

Here, C_1, C_2, C_3, C_4 are independent of y and s .

Proof Since λ is analytic in y , Taylor's Theorem allows us to expand λ as a zeroth order Taylor series (see, e.g., [21, pp. 12,13]) in the variables $\mathbf{y}_{\{j>s\}} := (y_j)_{j>s}$ about the point $\mathbf{0}$:

$$\lambda(\mathbf{y}) = \lambda(\mathbf{y}_s; \mathbf{0}) + \sum_{j>s} \int_0^1 \left(\frac{\partial \lambda}{\partial y_j} \right) (\mathbf{y}_s; t \mathbf{y}_{\{j>s\}}) y_j \, dt.$$

Noting that by definition $\lambda(\mathbf{y}_s; \mathbf{0}) = \lambda_s(\mathbf{y}_s)$, using the triangle inequality, the fact that $|y_j| \leq \frac{1}{2}$ and the upper bound (3.6), it follows that

$$|\lambda(\mathbf{y}) - \lambda_s(\mathbf{y}_s)| \leq \frac{\overline{\lambda_1}}{2} \sum_{j>s} \beta_j. \quad (4.5)$$

The eigenfunction is also analytic, and so similarly

$$\begin{aligned} \|u(\mathbf{y}) - u_s(\cdot, \mathbf{y}_s)\|_V &= \left\| \sum_{j>s} \int_0^1 \left(\frac{\partial u}{\partial y_j} \right) (\mathbf{y}_s; t \mathbf{y}_{\{j>s\}}) y_j \, dt \right\|_V \\ &\leq \sum_{j>s} \frac{1}{2} \int_0^1 \left\| \left(\frac{\partial u}{\partial y_j} \right) (\mathbf{y}_s; t \mathbf{y}_{\{j>s\}}) \right\|_V \, dt \leq \frac{\overline{u_1}}{2} \sum_{j>s} \beta_j, \end{aligned} \quad (4.6)$$

where we have used the upper bound (3.7).

In [26, Theorem 5.1] it was shown that under Assumption A1.3 the tail of the sum of the β_j is bounded above by

$$\sum_{j>s} \beta_j \leq \underbrace{\min \left(\frac{p}{1-p}, 1 \right) \|\boldsymbol{\beta}\|_{\ell^p} s^{-1/p+1}}_{:= C_{\text{trunc}}}, \quad (4.7)$$

which after substitution into (4.5) and (4.6) yields the two results for the strong error, (4.1) and (4.2), respectively.

For the weak error (4.3) we now use a k th order Taylor series expansion (see, e.g., [21, pp. 12,13]), then handle the Taylor sum and the remainder term separately.

First, we introduce some notation: let $k = \lceil 1/(1-p) \rceil$, and note that $1 < k < \infty$. Define $\mathcal{F}_s := \{\mathbf{0} \neq \mathbf{v} \in \mathcal{F} : v_j = 0 \text{ for all } j = 1, 2, \dots, s\}$, and for $\ell \in \mathbb{N}$ let $\mathcal{F}_{\ell,s} := \{\mathbf{v} \in \mathcal{F}_s : |\mathbf{v}| = \ell\}$.

The k th order Taylor series expansion of λ in the variables $\mathbf{y}_{\{j>s\}}$ is

$$\begin{aligned}\lambda(\mathbf{y}) &= \lambda(\mathbf{y}_s; \mathbf{0}) + \sum_{\ell=1}^k \sum_{\mathbf{v} \in \mathcal{F}_{\ell,s}} \frac{\mathbf{y}^\mathbf{v}}{\mathbf{v}!} (\partial^\mathbf{v} \lambda)(\mathbf{y}_s; \mathbf{0}) \\ &\quad + \sum_{\mathbf{v} \in \mathcal{F}_{k+1,s}} \frac{k+1}{\mathbf{v}!} \mathbf{y}^\mathbf{v} \int_0^1 (1-t)^k (\partial^\mathbf{v} \lambda)(\mathbf{y}_s; t \mathbf{y}_{\{j>s\}}) dt.\end{aligned}$$

Taking the expected value with respect to \mathbf{y} , by linearity we obtain

$$\begin{aligned}\mathbb{E}_\mathbf{y}[\lambda - \lambda_s] &= \sum_{\ell=1}^k \sum_{\mathbf{v} \in \mathcal{F}_{\ell,s}} \frac{1}{\mathbf{v}!} \mathbb{E}_\mathbf{y} [\mathbf{y}^\mathbf{v} (\partial^\mathbf{v} \lambda)(\mathbf{y}_s; \mathbf{0})] \\ &\quad + \sum_{\mathbf{v} \in \mathcal{F}_{k+1,s}} \frac{k+1}{\mathbf{v}!} \mathbb{E}_\mathbf{y} \left[\mathbf{y}^\mathbf{v} \int_0^1 (1-t)^k (\partial^\mathbf{v} \lambda)(\mathbf{y}_s; t \mathbf{y}_{\{j>s\}}) dt \right].\end{aligned}\quad (4.8)$$

Since each y_j is independent, for $\mathbf{v} \in \mathcal{F}_{\ell,s}$, we have

$$\mathbb{E}_\mathbf{y} [\mathbf{y}^\mathbf{v} (\partial^\mathbf{v} \lambda)(\mathbf{y}_s; \mathbf{0})] = \mathbb{E}_\mathbf{y} [(\partial^\mathbf{v} \lambda)(\mathbf{y}_s; \mathbf{0})] \prod_{j>s} \mathbb{E}_\mathbf{y}[y_j^{v_j}],$$

and because y_j has mean 0, any term in the first sum in (4.8) with at least one $v_j = 1$ is zero. This means that for the first term in (4.8) we only need to consider higher order derivatives ($\ell \geq 2$), and we can restrict the inner sum to $\mathbf{v} \in \mathcal{F}_{\ell,s}$ such that $v_j \neq 1$, giving

$$\begin{aligned}\mathbb{E}_\mathbf{y}[\lambda - \lambda_s] &= \sum_{\ell=2}^k \sum_{\substack{\mathbf{v} \in \mathcal{F}_{\ell,s} \\ v_j \neq 1 \forall j}} \frac{1}{\mathbf{v}!} \mathbb{E}_\mathbf{y} [\mathbf{y}^\mathbf{v} (\partial^\mathbf{v} \lambda)(\mathbf{y}_s; \mathbf{0})] \\ &\quad + \sum_{\mathbf{v} \in \mathcal{F}_{k+1,s}} \frac{k+1}{\mathbf{v}!} \mathbb{E}_\mathbf{y} \left[\mathbf{y}^\mathbf{v} \int_0^1 (1-t)^k (\partial^\mathbf{v} \lambda)(\mathbf{y}_s; t \mathbf{y}_{\{j>s\}}) dt \right].\end{aligned}$$

Taking the absolute value then using the triangle inequality, monotonicity of the expectation and that $|y_j| \leq 1/2$, we have the bound

$$\begin{aligned} |\mathbb{E}_y[\lambda - \lambda_s]| &\leq \sum_{\ell=2}^k \sum_{\substack{\nu \in \mathcal{F}_{\ell,s} \\ \nu_j \neq 1 \forall j}} \frac{1}{2^\ell \nu!} \mathbb{E}_y [(\partial^\nu \lambda)(\mathbf{y}_s; \mathbf{0})] \\ &+ \sum_{\nu \in \mathcal{F}_{k+1,s}} \frac{k+1}{2^{k+1} \nu!} \mathbb{E}_y \left[\int_0^1 (1-t)^k |(\partial^\nu \lambda)(\mathbf{y}_s; t\mathbf{y}_{\{j>s\}})| dt \right]. \end{aligned}$$

Bounding each derivative using (3.6), which is independent of \mathbf{y} , and then evaluating the remaining integral over t exactly gives

$$\begin{aligned} |\mathbb{E}_y[\lambda - \lambda_s]| &\leq \sum_{\ell=2}^k \sum_{\substack{\nu \in \mathcal{F}_{\ell,s} \\ \nu_j \neq 1 \forall j}} \frac{\bar{\lambda}(\ell!)^{1+\epsilon}}{2^\ell \nu!} \beta^\nu + \sum_{\nu \in \mathcal{F}_{k+1,s}} \frac{\bar{\lambda}((k+1)!)^{1+\epsilon}}{2^{k+1} \nu!} \beta^\nu \\ &\leq C_k \sum_{\ell=2}^k \sum_{\substack{\nu \in \mathcal{F}_{\ell,s} \\ \nu_j \neq 1 \forall j}} \beta^\nu + C_{k+1} \sum_{\nu \in \mathcal{F}_{k+1,s}} \beta^\nu, \end{aligned} \quad (4.9)$$

where

$$C_k = \bar{\lambda} \max_{\substack{\nu \in \mathcal{F} \\ |\nu| \leq k}} \frac{(|\nu|!)^{1+\epsilon}}{2^{|\nu|} \nu!} < \infty$$

is independent of s , but depends on p through k , and we have again dropped the subscript 1 for the upper bound (see (2.14)) on the minimal eigenvalue, $\bar{\lambda} := \bar{\lambda}_1$.

For the first sum in (4.9), since the order $|\nu|$ satisfies $|\nu| \geq \|\nu\|_{\ell^\infty}$ and every β_j is positive, we can add extra terms to the sum to obtain the bound

$$\sum_{\ell=2}^k \sum_{\substack{\nu \in \mathcal{F}_{\ell,s} \\ \nu_j \neq 1 \forall j}} \beta^\nu = \sum_{\substack{\nu \in \mathcal{F}_s \\ |\nu| \leq k \\ \nu_j \neq 1 \forall j}} \beta^\nu \leq \sum_{\substack{\nu \in \mathcal{F}_s \\ \|\nu\|_{\ell^\infty} \leq k \\ \nu_j \neq 1 \forall j}} \beta^\nu.$$

Then, as in [17] we can write the sum on the right as the following product

$$\sum_{\substack{\nu \in \mathcal{F}_s \\ \|\nu\|_{\ell^\infty} \leq k \\ \nu_j \neq 1 \forall j}} \beta^\nu = -1 + \prod_{j>s} \left(1 + \sum_{\ell=2}^k \beta_j^\ell \right) = -1 + \prod_{j>s} \left(1 + \left(\frac{1 - \beta_j^{k-1}}{1 - \beta_j} \right) \beta_j^2 \right),$$

where in the last step we have used the formula for the sum of a geometric series.

In order to simplify the product above, we define the sequence $\tilde{\beta}$ by

$$\tilde{\beta}_j := \begin{cases} \beta_j^2, & \text{for } j = 1, 2, \dots, s, \\ \left(\frac{1 - \beta_j^{k-1}}{1 - \beta_j} \right) \beta_j^2, & \text{for } j > s, \end{cases}$$

which, because the sequence β is assumed to be decreasing, is well defined for s sufficiently large such that $\beta_s \leq 1/2$. Further, since $\tilde{\beta}_j \leq \beta_j^2/(1 - \beta_s) \leq 2\beta_j^2$ for all $j \in \mathbb{N}$, it follows that $\tilde{\beta} \in \ell^{p/2}$ and $\|\tilde{\beta}\|_{\ell^{p/2}}$ can be bounded from above independently of s .

Then, using the inequalities $\ln(1 + x) \leq x$ and $-1 + e^x \leq xe^x$, we can bound the first sum in (4.9) by

$$\begin{aligned} \sum_{v=2}^k \sum_{\substack{v \in \mathcal{F}_{t,s} \\ v_j \neq 1 \forall j}} \beta^v &\leq -1 + \prod_{j>s} (1 + \tilde{\beta}_j) = -1 + \exp \left(\sum_{j>s} \ln(1 + \tilde{\beta}_j) \right) \leq \exp \left(\sum_{j>s} \tilde{\beta}_j \right) \sum_{j>s} \tilde{\beta}_j \\ &\leq \exp (\|\tilde{\beta}\|_{\ell^1}) \min \left(\frac{p}{2-p}, 1 \right) \|\tilde{\beta}\|_{\ell^{p/2}} s^{-2/p+1}, \end{aligned} \quad (4.10)$$

where we have also used (4.7) for the sequence $\tilde{\beta} \in \ell^{p/2}$.

For the second sum in (4.9), since $\binom{k+1}{v} = \frac{(k+1)!}{v!} \geq 1$, then using (4.7) and the definition of k we have

$$\begin{aligned} \sum_{v \in \mathcal{F}_{k+1,s}} \beta^v &\leq \sum_{v \in \mathcal{F}_{k+1,s}} \binom{k+1}{v} \beta^v = \left(\sum_{j>s} \beta_j \right)^{k+1} \leq C_{\text{trunc}}^{k+1} s^{(k+1)(-1/p+1)} \\ &\leq C_{\text{trunc}}^{k+1} s^{(1/(1-p)+1)(-1/p+1)} = C_{\text{trunc}}^{k+1} s^{-2/p+1}. \end{aligned} \quad (4.11)$$

The result (4.3) for the weak error of the eigenvalue is then obtained by substituting (4.10) and (4.11) into (4.9).

As for the proof of the strong error (4.2), we can also expand u as a k th order Taylor series, and then use the same argument to prove the bound (4.4) for the weak error of the eigenfunction. \square

4.2 QMC error

Given the bounds in Lemma 3.4 on the mixed derivatives of the minimal eigenpair we now obtain an upper bound of the root-mean-square error of the QMC approximation of the truncated problem.

Theorem 4.2 *Let $N \in \mathbb{N}$ be prime, $\mathcal{G} \in V^*$ and suppose that Assumption A1 holds. Then the root-mean-square errors of the CBC-generated randomly shifted lattice rule approximations of $\mathbb{E}_y [\lambda_s]$ and $\mathbb{E}_y [\mathcal{G}(u_s)]$ are bounded by*

$$\sqrt{\mathbb{E}_\Delta \left[|\mathbb{E}_y [\lambda_s] - Q_{N,s} \lambda_s|^2 \right]} \leq C_{1,\alpha} N^{-\alpha}, \quad \text{and} \quad (4.12)$$

$$\sqrt{\mathbb{E}_\Delta \left[|\mathbb{E}_y [\mathcal{G}(u_s)] - Q_{N,s} \mathcal{G}(u_s)|^2 \right]} \leq C_{2,\alpha} N^{-\alpha}, \quad (4.13)$$

where

$$\alpha = \begin{cases} 1 - \delta, & \text{for arbitrary } \delta \in (0, \frac{1}{2}), \text{ if } p \in (0, \frac{2}{3}], \\ \frac{1}{p} - \frac{1}{2} & \text{if } p \in (\frac{2}{3}, 1), \end{cases} \quad (4.14)$$

and the constants $C_{1,\alpha}$ and $C_{2,\alpha}$ are independent of s .

Proof Since the estimates from Lemma 3.4 are independent of y they can be used to bound the norm (squared) of λ_s in $\mathcal{W}_{s,\gamma}$. By (3.6) we obtain

$$\|\lambda_s\|_{s,\gamma}^2 \leq \bar{\lambda}^2 \sum_{u \subseteq \{1:s\}} \frac{\Lambda_u^2}{\gamma_u}, \quad \Lambda_u := (|u|!)^{1+\epsilon} \prod_{j \in u} \beta_j,$$

with weights γ and $\epsilon \in (0, 1)$ as yet unspecified. Then, using (2.37) the mean-square error of the lattice rule approximation is bounded above by

$$\mathbb{E}_\Delta \left[|\mathbb{E}_y [\lambda_s] - Q_{N,s} \lambda_s|^2 \right] \leq C_{s,\gamma,\eta} \varphi(N)^{-\frac{1}{\eta}}, \quad (4.15)$$

where

$$C_{s,\gamma,\eta} := \bar{\lambda}^2 \left(\sum_{\emptyset \neq u \subseteq \{1:s\}} \gamma_u^\eta \rho(\eta)^{|u|} \right)^{\frac{1}{\eta}} \left(\sum_{u \subseteq \{1:s\}} \frac{\Lambda_u^2}{\gamma_u} \right)$$

We now choose the weight parameters such that $C_{s,\gamma,\eta}$ can be bounded independently of s . From [26, Lemma 6.2] the choice of weights that minimise $C_{s,\gamma,\eta}$ are

$$\gamma_u(\eta) = \left(\frac{\Lambda_u^2}{\rho(\eta)^{|u|}} \right)^{\frac{1}{1+\eta}}, \quad (4.16)$$

which are of POD (product and order-dependent) form. With these weights it follows that $C_{s,\gamma,\eta} \leq \bar{\lambda}^2 S_{s,\eta}^{(1+\eta)/\eta}$, where

$$S_{s,\eta} := \sum_{u \subseteq \{1:s\}} \left(\Lambda_u^{2\eta} \rho(\eta)^{|u|} \right)^{\frac{1}{1+\eta}},$$

so we must show that the sum $S_{s,\eta}$ can be bounded independently of s . Let

$$q := \frac{2\eta(1+\epsilon)}{1+\eta} \quad \text{and} \quad \alpha_j := \left(\rho(\eta)(\beta_j)^{2\eta} \right)^{\frac{1}{1+\eta}} \quad \text{for all } j \in \mathbb{N},$$

so that

$$S_{s,\eta} = \sum_{\ell=0}^s (\ell!)^q \sum_{\substack{\mathbf{u} \subseteq \{1:s\} \\ |\mathbf{u}|=\ell}} \prod_{j \in \mathbf{u}} \alpha_j \leq \sum_{\ell=0}^s (\ell!)^{q-1} \left(\sum_{j=1}^s \alpha_j \right)^\ell < \infty,$$

which holds by the ratio test provided that $q < 1$ and $\sum_{j=1}^\infty \alpha_j < \infty$. Under Assumption A1.3, we therefore require that

$$\frac{2\eta(1+\epsilon)}{1+\eta} < 1 \iff \epsilon < \frac{1-\eta}{2\eta} \quad \text{and} \quad \frac{2\eta}{1+\eta} \leq p \iff \eta \geq \frac{p}{2-p}.$$

To balance these conditions with the requirement that $\eta \in (\frac{1}{2}, 1]$, we choose a different η depending on the decay rate p and then choose $\epsilon := (1-\eta)/(4\eta)$. Note that $\eta = 1$, equivalently $p = 1$, has to be excluded to ensure that $\epsilon > 0$.

For $p \in (0, \frac{2}{3}]$, we have $\frac{p}{2-p} \leq \frac{1}{2}$ so there is no further restriction on η and we take $\eta := \frac{1}{2(1-\delta)}$ for arbitrary $\delta \in (0, \frac{1}{2})$. However, for $p \in (\frac{2}{3}, 1)$ the value of η is restricted and we take it as small as possible, namely, $\eta := \frac{p}{2-p}$. Substituting these choices of η into (4.15) and taking N to be prime (for simplicity) yields the result (4.12).

The error bound (4.13) follows in the same way, after observing that the norm of $\mathcal{G}(u_s)$ can be bounded using (3.7)

$$\begin{aligned} \|\mathcal{G}(u_s)\|_{s,\gamma}^2 &\leq \sum_{\mathbf{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathbf{u}}} \int_{[0,1]^{|\mathbf{u}|}} \left(\int_{[0,1]^{s-|\mathbf{u}|}} \|\mathcal{G}\|_{V^*} \left\| \frac{\partial^{|\mathbf{u}|} u_s}{\partial \mathbf{y}_{\mathbf{u}}}(\cdot, \mathbf{y}_s) \right\|_V d\mathbf{y}_{-u} \right)^2 d\mathbf{y}_{\mathbf{u}} \\ &\leq \|\mathcal{G}\|_{V^*}^2 \bar{u}^2 \sum_{\mathbf{u} \subseteq \{1:s\}} \frac{\Lambda_{\mathbf{u}}^2}{\gamma_{\mathbf{u}}}. \end{aligned}$$

This completes the proof. \square

Remark 4.3 The main ingredients in this proof are the bounds on the derivatives of the eigenvalue, which are needed to show that $\lambda_s \in \mathcal{W}_{s,\gamma}$. As was stated in Remark 3.5, these bounds also hold for the derivatives of $\lambda_{s,h}$. Thus, for h sufficiently small (see (3.3)) the same error bound holds for the QMC error of the FE error approximation $\lambda_{s,h}$, but with the constants possibly depending on h .

4.3 Total error

Using the triangle inequality to give, the mean-square error of the combined truncation-FE-QMC approximation of the expected value of λ_1 can be bounded above by

$$\begin{aligned} \mathbb{E}_\Delta \left[|\mathbb{E}_y[\lambda] - Q_{N,s} \lambda_{s,h}|^2 \right] &\leq C \left(\mathbb{E}_\Delta \left[|\mathbb{E}_y[\lambda - \lambda_s]|^2 \right] \right. \\ &\quad \left. + \mathbb{E}_\Delta \left[|\mathbb{E}_y[\lambda_s] - Q_{N,s} \lambda_s|^2 \right] + \mathbb{E}_\Delta \left[|Q_{N,s}(\lambda_s - \lambda_{s,h})|^2 \right] \right), \end{aligned} \quad (4.17)$$

for $C > 0$. Here we have conveniently split the total error into three separate errors: one each for the truncation, QMC and FE errors, respectively. Note that there are different ways of splitting the total error, but we have chosen the above technique because now the second term is the QMC error for the actual eigenvalue and not the FE approximation $\lambda_{s,h}$. This is important because it means that we do not need specific bounds on the parametric regularity of the FE eigenvalue.

The terms in the upper bound on the mean-square error in (4.17) can be bounded using (4.3), (4.12) and (2.33), respectively, leading to the following theorem. A similar splitting argument, using (4.2), (2.34) and (4.13) instead, gives a bound on the error of the approximation for functionals of the corresponding eigenfunction.

Theorem 4.4 *Let Assumption A1 hold, $h > 0$ be sufficiently small, $s \in \mathbb{N}$, $N \in \mathbb{N}$ be prime and let $z \in \mathbb{N}^s$ be a generating vector constructed using the CBC algorithm with weights given by (4.16). Then the root-mean-square error, with respect to the random shift $\Delta \in [0, 1]^s$, of our truncation-FE-QMC approximation of the mean of the minimal eigenvalue λ is bounded by*

$$\sqrt{\mathbb{E}_\Delta \left[|\mathbb{E}_y[\lambda] - Q_{N,s}\lambda_{s,h}|^2 \right]} \leq C_1 \left(h^2 + s^{-2/p+1} + N^{-\alpha} \right). \quad (4.18)$$

For any functional $\mathcal{G} \in H^{-1+t}(D)$ applied to the corresponding eigenfunction u , with $t \in [0, 1]$, the truncation-FE-QMC approximation of its mean is bounded by

$$\sqrt{\mathbb{E}_\Delta \left[|\mathbb{E}_y[\mathcal{G}(u)] - Q_{N,s}\mathcal{G}(u_{s,h})|^2 \right]} \leq C_2 \left(h^{1+t} + s^{-2/p+1} + N^{-\alpha} \right). \quad (4.19)$$

Where

$$\alpha = \begin{cases} 1 - \delta, & \text{for arbitrary } \delta \in (0, \frac{1}{2}), \text{ if } p \in (0, \frac{2}{3}], \\ \frac{1}{p} - \frac{1}{2} & \text{if } p \in (\frac{2}{3}, 1), \end{cases}$$

and the constants $C_1, C_2 > 0$ are independent of s, h and N .

Comparing this with the corresponding result for the elliptic source problem [26, Theorem 8.1], observe that we obtain the exact same rate of convergence in N for all $p \in (0, 1)$. The only exception is that our results do not hold when $p = 1$, whereas [26, Theorem 8.1] presents a result when $p = 1$. However, for that case they do require an additional assumption (see [26, equation (6.5)]). For the truncation error, we obtain the same convergence rate as in [17], which improved the rate in [26, Theorem 8.1] by one order. To compare the two FE convergence rates recall that the number of degrees of freedom in the FE grid is $M_h = \mathcal{O}(h^{-d})$. Letting $\mathcal{G} \in H^{-1+t}(D)$, note that the correct comparison is the case when the source term belongs to $L^2(D)$, so that in [26, Theorem 8.1] $\tau = 1 + t$. In this case the FE error convergence rate from [26, Theorem 8.1] for the linear functional \mathcal{G} of the solution to the source problem is $M_h^{-\tau/d} = \mathcal{O}(h^{1+t})$, which is exactly the rate in (4.19) for the eigenfunction.

5 Numerical results

Now we present numerical results on the performance of our truncation-FE-QMC algorithm in approximating the expected value of the smallest eigenvalue of two different eigenvalue problems of the form (1.1). In both cases, the stochastic coefficients a and b are composed of scaled trigonometric functions, which can be seen as artificial Karhunen–Loève (KL) expansions. We will focus on the question of whether the error matches the theoretical estimate from Theorem 4.4, and so we will study different scalings of the basis functions a_j, b_j in the coefficients, which will correspond to different values of the decay parameter p . All computations were performed on the Katana cluster at UNSW Sydney.

In order to estimate the quadrature error, we conduct a small number R of different approximations based on independently and identically distributed random shifts. We label the approximation generated by the r th random shift $\Delta^{(r)}$ as $Q_{N,s}^{(r)}\lambda_{s,h}$, and the final approximation is taken to be the average over the R independent approximations:

$$\widehat{Q}_{R,N,s}\lambda_{s,h} := \frac{1}{R} \sum_{r=1}^R Q_{N,s}^{(r)}\lambda_{s,h}.$$

In this way, we obtain an unbiased estimate of the integral (of $\lambda_{s,h}$) and the sample standard deviation over the different shifts gives an estimate of the quadrature component of the RMS error

$$\sqrt{\mathbb{E}_\Delta \left[|\mathbb{E}_y[\lambda_{s,h}] - \widehat{Q}_{N,s}\lambda_{s,h}|^2 \right]} \approx \sqrt{\frac{1}{R(R-1)} \sum_{r=1}^R \left(Q_{N,s}^{(r)}\lambda_{s,h} - \widehat{Q}_{R,N,s}\lambda_{s,h} \right)^2}. \quad (5.1)$$

Note that averaging over R shifts increases the total number of function evaluations by a factor of R , and hence, one can expect the error to correspondingly decrease by a factor of $1/\sqrt{R}$ (in line with the Monte Carlo rate).

The smallest eigenvalue of each FE system is approximated using the `eigs` function in Matlab, which in turn runs a Krylov–Schur algorithm using the ARPACK library. We set the tolerance for the accuracy of this eigensolver to be 10^{-14} , so as to ensure that the numerical errors incurred in computing the FE eigenvalues are negligible compared to the overall approximation error.

In practice, we cannot compute the optimal function space weights γ_u according to the formula (4.16) from the proof of Theorem 4.2, because (4.16) depends on the sequence β , and C_β (3.9) contains factors that cannot be computed explicitly. Instead we choose the weights so that the product components decay at the same rate as in the optimal formula (4.16). As such, in our numerical experiments for $u \subset \mathbb{N}$ we set the function space weight γ_u to be

$$\gamma_u = |u|! \prod_{j \in u} \left(\max \{ \|a_j\|_{L^\infty(D)}, \|b_j\|_{L^\infty(D)} \} \right)^\eta, \quad (5.2)$$

where $\eta = 4/3$ if $p \in (0, 2/3]$ and $\eta = 2 - p$ if $p \in (2/3, 1)$. Note that η in (5.2) is not the same as in (4.16).

5.1 Problem 1

In our first simple example, we consider an eigenvalue problem (1.1) on the domain $D = (0, 1)^2$ where the only non-trivial coefficient is $a(\mathbf{y})$ in the second-order term. Explicitly, the coefficient $a(\mathbf{y})$ is given as in (1.3) (with a_j defined below) but $b(\mathbf{y}) \equiv 0$ and $c \equiv 1$. For some decay $q \geq 4/3$, the basis functions for the coefficient $a(\mathbf{y})$ are defined to be

$$\begin{aligned} a_0 &\equiv 1, \\ a_j(\mathbf{x}) &= \frac{1}{1 + (j\pi)^q} \sin(j\pi x_1) \sin((j+1)\pi x_2), \quad \text{for } \mathbf{x} = (x_1, x_2) \in (0, 1)^2. \end{aligned} \tag{5.3}$$

Clearly, for all $j \in \mathbb{N}$ we have that

$$\|a_j\|_{L^\infty(D)} = \frac{1}{1 + (j\pi)^q} < \frac{1}{(j\pi)^q},$$

and hence $\sum_{j=1}^{\infty} \|a_j\|_{L^\infty(D)} < \zeta(q)/\pi^q$, where again ζ is the Riemann zeta function. It follows that the coefficient is bounded above and below as required with

$$a_{\min} = 1 - \frac{\zeta(q)}{\pi^q} \quad \text{and} \quad a_{\max} = 1 + \frac{\zeta(q)}{\pi^q}.$$

Similarly, the parameter q determines the rate of decay of the norms of the basis functions, and in turn we can take p in Assumption A1.3 to satisfy $p \in (1/q, 1)$. We also ran the case when the coefficient b is nonzero and of a similar form to (5.3). The results are almost exactly the same as this example, and so have not been included.

In our numerical experiments for this problem we consider $q = 4/3, 2, 3$, and vary the approximation parameters as follows. The truncation dimensions tested are given by $s = 2, 4, 8, \dots, 256$; we use uniform triangular FE meshes with $h = 1/4, 1/8, 1/16, \dots, 1/1024$; and the number of quadrature points is given by $N = 31, 61, 127, 251, 503, 997, 1999, 4001, 8009, 16001$.

Regarding the different QMC convergence rates to expect, for $q = 4/3$ we have that the sequence β is p -summable for $p > 3/4$, whereas for the faster decays of $q = 2, 3$ we have that β is summable with exponent $p > 1/2$ and $p > 1/3$, respectively. Based on these restrictions on p , for each N we construct a generating vector by the CBC algorithm using weights given by (5.2) with $\eta = 2 - 1/q = 5/4$ for $q = 4/3$, and $\eta = 4/3$ for $q = 2, 3$.

First, we study the truncation error by varying s , while keeping $h = 1/512$ and $N = 8009$ fixed (using a single fixed shift for all values of s). The results are given in Fig. 1. The errors are estimated by comparing each result with a fine solution with truncation dimension 512. Theorem 4.4 predicts that the truncation errors converge

Fig. 1 Dimension truncation error estimate for $q = 4/3, 2, 3$
($p \approx 3/4, 1/2, 1/3$)

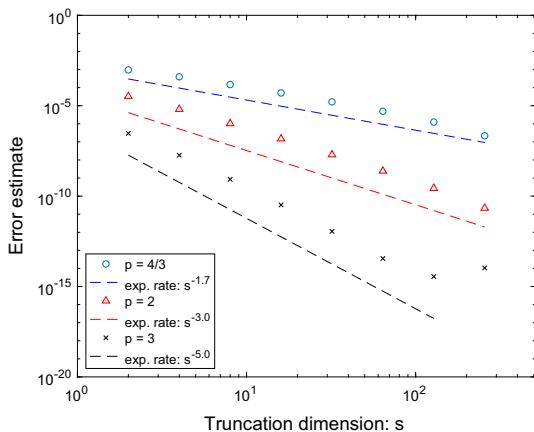
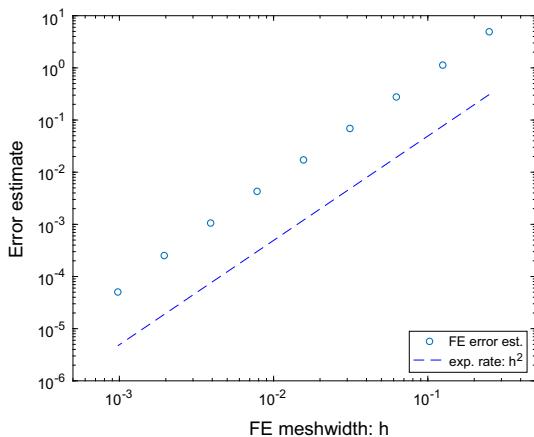


Fig. 2 FE error estimate for
 $q = 4/3, 2, 3$
($p \approx 3/4, 1/2, 1/3$)



like $s^{-5/3}$, s^{-3} and s^{-5} for $q = 4/3, 2, 3$, respectively. As a guide, these expected rates are given by the dashed lines in Fig. 1. Observe that for all cases the estimated truncation errors closely follow the expected convergence.

In Fig. 2 we present results for the FE error convergence for $q = 4/3$ ($p \approx 3/4$). Again, to isolate the FE component we vary h , but fix $s = 128$ and $N = 8009$ (using a single fixed shift for all values of h). Then we estimate the errors by comparing with a fine solution that uses a meshwidth of $1/2048$. Theorem 4.4 predicts that the FE error converges like h^2 , which is clearly observed. The other cases, $q = 2, 3$, both exhibit very similar errors and so have not been included.

To study the quadrature component of the error, we fix $s = 256$, $h = 1/256$ and use $R = 8$ random shifts to estimate the RMS error (5.1). Figure 3 plots the estimated RMS quadrature error of our QMC approximation along with a Monte Carlo (MC) approximation for comparison, for $q = 4/3$ (left) and $q = 2$ (right). To fairly compare with the MC results, for each N we present the estimated RMS error of a single randomly shifted lattice rule. That is, we scale the RMS error estimate in (5.1) by \sqrt{R} so as to remove the extra $1/\sqrt{R}$ factor gained by random shifting. In Fig. 3

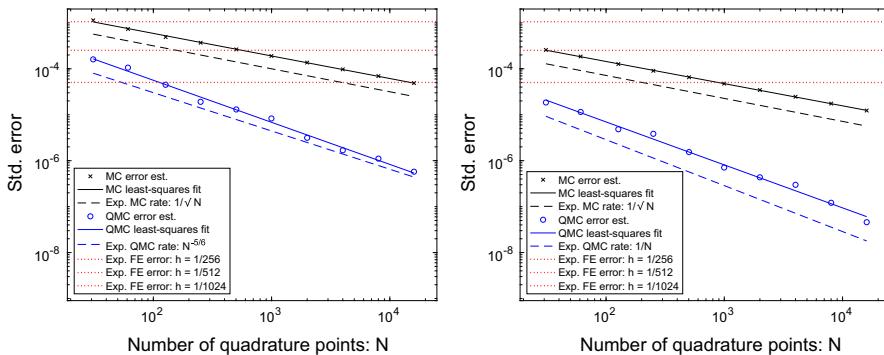


Fig. 3 QMC and MC convergence for $q = 4/3$, $p \approx 3/4$ (left) and $q = 2$, $p \approx 1/2$ (right)

the circle data points (crosses for MC) represent the estimated RMS errors, the dashed lines portray the expected convergence rates and the solid lines are a least-squares fit. As a guide, we have plotted the FE error one can expect for $h = 2^{-8}, 2^{-9}, 2^{-10}$ as three red dotted lines (we use the results used to generate Fig. 2), note that the vertical heights of the lines decrease as h decreases.

From Theorem 4.4, for $q = 4/3$ (because $p > 2/3$) we are in the regime where the convergence is limited and so expect a rate of around $N^{-5/6}$. However, for $q = 2$ the rate is not restricted and we expect QMC convergence close to N^{-1} . Observe that the least-squares fits match very closely to the expected rates. Also, as is expected for a problem of this smoothness QMC significantly outperforms the MC approximations, which decay at the anticipated rate of $1/\sqrt{N}$. Comparing with the expected FE errors (the red dotted lines), we can see that for this example the dominant component of the error is the FE error, this is to be expected and has been observed in source problems also. Despite this, we have continued with QMC approximations that have errors below the FE error to demonstrate that the asymptotic convergence rates predicted by Theorem 4.4 are achieved in practice. For the truncation error, taking $s = 256$ results in errors of approximately 10^{-7} and 10^{-10} for $q = 4/3$ and 2, respectively, which are at least an order of magnitude less than the smallest quadrature errors in each case.

For completeness, the computed values of the RMS error estimates for the shift-averaged estimate $\widehat{Q}_{R,N,s}\lambda_{s,h}$ and the convergence rates for $q = 4/3, 2$ and 3 are given below in Table 1. We use the notation “e” to denote the base 10 exponent. The least-squares computed rates are -0.878 , -0.992 and -1.01 for $q = 4/3, 2, 3$, respectively, which are very close to the expected rates of $-5/6 \approx -0.83$, -1 , and -1 from the theory.

5.2 Problem 2

For our second example we consider more complicated coefficients, chosen to represent a problem where the physical domain is composed of different materials, e.g., in a simple model of a nuclear reactor the domain is composed of fuel rods surrounded by coolant (a gas/water mixture). In particular, we allow the right hand side coefficient

Table 1 QMC RMS error estimates for $q = 4/3, 2, 3$

N	$R \times N$	RMS error estimate of $\widehat{Q}_{R,N,s} \lambda_{s,h}$		
		$q = 4/3$ ($p \approx 3/4$)	$q = 2$ ($p \approx 1/2$)	$q = 3$ ($p \approx 1/3$)
251	2008	6.8e-6	1.4e-6	4.8e-8
503	4024	4.6e-6	5.4e-7	1.8e-8
997	7976	2.9e-6	2.5e-7	7.3e-9
1999	15,992	1.1e-6	1.5e-7	5.8e-9
4001	32,008	6.0e-7	1.1e-7	2.5e-9
8009	64,072	3.9e-7	4.3e-8	1.1e-9
16,001	128,008	2.0e-7	1.6e-8	6.4e-10
Estimated rate		-0.878	-0.992	-1.01

c to be stochastic, and zero on parts of the domain. Although this problem does not satisfy Assumption A1, we have included it to illustrate that our method also works well for a larger class of eigenvalue problems than what we were able to analyse theoretically.

In order to focus on the behaviour of the QMC error, in this numerical experiment we fix the truncation dimension at $s = 100$ and FE meshwidth at $h = 1/256$, and we vary the number of QMC points N . Also, we set the number of random shifts to be $R = 8$.

Again the domain is $D = (0, 1)^2$ and we define $\mathbf{x} := (x_1, x_2)$, but now the coefficients have different basis expansions for two different components of the domain depicted in Fig. 4, where D is separated into the union of four islands (in grey and labelled D_f) and the area around the islands $D \setminus D_f$ (in white). We define D_f by

$$D_f := [\frac{1}{8}, \frac{3}{8}]^2 \cup [\frac{5}{8}, \frac{7}{8}]^2 \cup [\frac{1}{8}, \frac{3}{8}] \times [\frac{5}{8}, \frac{7}{8}] \cup [\frac{5}{8}, \frac{7}{8}] \times [\frac{1}{8}, \frac{3}{8}].$$

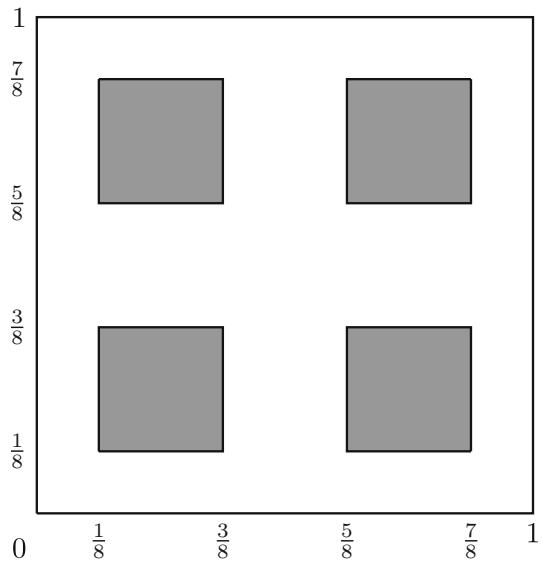
Since we use a uniform triangular FE mesh with $h = 1/256$ the FE triangulation aligns with the boundaries of the components D_f .

Again, we would like the coefficients to have the form of an artificial KL expansion, but we would also like the flexibility to be able to specify different decays and scalings for each coefficient on the different components. To achieve this, for $k \in \mathbb{N}$, we define the following sequences of trigonometric functions

$$w_k(q; \mathbf{x}) = \begin{cases} \frac{1}{1 + (k\pi)^q} \sin(8k\pi x_1) \sin(8(k+1)\pi x_2) & \text{for } \mathbf{x} \in D_f, \\ 0 & \text{for } \mathbf{x} \in D \setminus D_f, \text{ and} \end{cases}$$

$$w'_k(q; \mathbf{x}) = \begin{cases} 0 & \text{for } \mathbf{x} \in D_f, \\ \frac{1}{1 + (k\pi)^q} \sin(8k\pi x_1) \sin(8(k+1)\pi x_2) & \text{for } \mathbf{x} \in D \setminus D_f, \end{cases}$$

Fig. 4 Domain D with four islands forming D_f (in grey)



which have wavelengths chosen such that their zeros align with the boundaries of D_f . Now we use these functions to define the basis functions of the coefficients by

$$\begin{aligned} a_0(\mathbf{x}) &= \begin{cases} \sigma_{\text{diff}} := 0.01 & \text{for } \mathbf{x} \in D_f, \\ \sigma'_{\text{diff}} := 0.011 & \text{for } \mathbf{x} \in D \setminus D_f, \end{cases} & a_j(\mathbf{x}) &= \begin{cases} \sigma_{\text{diff}} w_{(j+1)/2}(q_a; \mathbf{x}) & \text{for } j \text{ odd}, \\ \sigma'_{\text{diff}} w'_{j/2}(q'_a; \mathbf{x}) & \text{for } j \text{ even}, \end{cases} \\ b_0(\mathbf{x}) &= \begin{cases} \sigma_{\text{abs}} := 2 & \text{for } \mathbf{x} \in D_f, \\ \sigma'_{\text{abs}} := 0.3 & \text{for } \mathbf{x} \in D \setminus D_f, \end{cases} & b_j(\mathbf{x}) &= \begin{cases} \sigma_{\text{abs}} w_{(j+1)/2}(q_b; \mathbf{x}) & \text{for } j \text{ odd}, \\ \sigma'_{\text{abs}} w'_{j/2}(q'_b; \mathbf{x}) & \text{for } j \text{ even}, \end{cases} \end{aligned}$$

where the parameters $q_a, q'_a, q_b, q'_b \geq 4/3$ give the different decays of the coefficients on the different areas of the domain.

The coefficients a , b , and c represent, respectively, diffusion, absorption and fission in the reactor. Motivated by reactors where fission only occurs in fuel rods, we assume that c vanishes outside D_f . Furthermore, we also let c be stochastic and take the same form as a and b . We define

$$c_0(\mathbf{y}) = \begin{cases} \sigma_{\text{fiss}} := 2.5 & \text{for } \mathbf{x} \in D_f, \\ 0 & \text{for } \mathbf{x} \in D \setminus D_f, \end{cases} \quad c_j(\mathbf{y}) = \begin{cases} \sigma_{\text{fiss}} w_{(j+1)/2}(2; \mathbf{x}) & \text{for } j \text{ odd}, \\ 0 & \text{for } j \text{ even}, \end{cases}$$

and then

$$c(\mathbf{x}, \mathbf{y}) = \begin{cases} \sigma_{\text{fiss}} + \sum_{j=1}^{\infty} y_{(j+1)/2} c_j(\mathbf{x}) & \text{for } \mathbf{x} \in D_f, \\ 0 & \text{for } \mathbf{x} \in D \setminus D_f. \end{cases}$$

The factors σ_{diff} , σ'_{diff} , σ_{abs} , σ'_{abs} , and σ_{fiss} are chosen so that the mean of a , b and c on the different components of the domain correspond to physically relevant values for the respective cross-sections for a nuclear reactor (see [37]).

Also, notice that the coefficients a , b and c will be correlated. The practical motivation for this is that the randomness at a point is generated by uncertainty in the material properties at that point, and thus will affect each coefficient in the same manner. However, for $\mathbf{x} \in D_f$ and $\mathbf{x}' \in D \setminus D_f$ the values of the coefficients $a(\mathbf{x}, \mathbf{y})$, $b(\mathbf{x}, \mathbf{y})$, $c(\mathbf{x}, \mathbf{y})$ are not correlated with any of $a(\mathbf{x}', \mathbf{y})$, $b(\mathbf{x}', \mathbf{y})$, $c(\mathbf{x}', \mathbf{y})$.

We have that

$$\|a_j\|_{L^\infty(D)} = \begin{cases} \frac{\sigma_{\text{diff}}}{1 + (\frac{j+1}{2}\pi)^{q_a}} & \text{if } j \text{ is odd,} \\ \frac{\sigma'_{\text{diff}}}{1 + (\frac{j}{2}\pi)^{q'_a}} & \text{if } j \text{ is even,} \end{cases} \quad \text{and} \quad \|b_j\|_{L^\infty(D)} = \begin{cases} \frac{\sigma_{\text{abs}}}{1 + (\frac{j+1}{2}\pi)^{q_b}} & \text{if } j \text{ is odd,} \\ \frac{\sigma'_{\text{abs}}}{1 + (\frac{j}{2}\pi)^{q'_b}} & \text{if } j \text{ is even.} \end{cases}$$

Letting $q := \min(q_a, q'_a, q_b, q'_b)$, a quick calculation gives that

$$\max(\|a_j\|_{L^\infty(D)}, \|b_j\|_{L^\infty(D)}) < \frac{\max(\sigma_{\text{diff}}, \sigma'_{\text{diff}}, \sigma_{\text{abs}}, \sigma'_{\text{abs}})}{(\pi j/2)^q} = \frac{2}{(\pi j/2)^q}.$$

Hence, since $q \geq 4/3$ the coefficients are bounded from above and below independently of \mathbf{y} , and we can take

$$a_{\min} = 0.01 \left(1 - \zeta(q) \left(\frac{2}{\pi}\right)^q\right) \quad \text{and} \quad a_{\max} = 2.5 \left(1 + \zeta(q) \left(\frac{2}{\pi}\right)^q\right).$$

Again, Assumption A1.3 holds for $p \in (1/q, 1)$, and so for this example the convergence rate will be determined by the minimum of the four rates q_a, q'_a, q_b, q'_b .

For this example, we also provide results for a linear functional applied to $u_1(\mathbf{y})$. We define $\mathcal{G} \in V^*$ to be the linear functional that computes the average neutron flux over one of the islands $D_{f,1} := [\frac{1}{8}, \frac{3}{8}]^2$, which is given by

$$\mathcal{G}(u_1(\mathbf{y})) := \frac{1}{|D_{f,1}|} \int_{D_{f,1}} u_1(\mathbf{x}, \mathbf{y}) \, d\mathbf{x}.$$

The results for this example for different combinations of values $q_a, q'_a, q_b, q'_b \in \{4/3, 2\}$ are presented in Table 2. Each column corresponds to a different choice of decays, and presents results for both the eigenvalue and \mathcal{G} applied to the eigenfunction. We present RMS error estimates for increasing N , followed by the estimated convergence rate. Recall that $R = 8$ is the number of random shifts used.

Even though this example does not satisfy the conditions of our theory, we still obtain good convergence rates for the QMC error. For the first two columns of Table 2, at least one decay has the value $4/3$, hence, we expect summability with $p \approx 3/4$ and a convergence rate of $-5/6 \approx -0.83$. When $q_a = q'_a = q_b = q'_b = 4/3$ (first column in Table 2), the eigenvalue error decays slower than $N^{-5/6}$ but the eigenfunction error

Table 2 Quadrature results for the approximation of $\mathbb{E}_y[\lambda_1]$ and $\mathbb{E}_y[\mathcal{G}(u_1)]$ for different decays

$q_a, q'_a, q_b, q'_b =$	N	$R \times N$	RMS error estimate			
			4/3, 4/3, 4/3, 4/3	2, 4/3, 2, 4/3	2, 2, 2, 2	$\mathcal{G}(u_1)$
251	2008	2.4e-7	5.1e-6	6.7e-7	5.3e-6	6.4e-7
503	4024	9.7e-8	2.8e-6	3.6e-7	3.0e-6	4.5e-7
997	7976	9.1e-8	1.0e-6	2.9e-7	8.4e-7	2.7e-7
1999	15, 992	3.9e-8	6.9e-7	1.4e-7	6.9e-7	1.3e-7
4001	32, 008	2.0e-8	3.8e-7	4.2e-8	3.6e-7	3.8e-8
8009	64, 072	1.4e-8	1.9e-7	5.1e-8	1.8e-7	4.8e-8
16,001	128, 008	1.1e-8	7.4e-8	1.7e-8	7.1e-8	9.0e-9
Estimated rate		-0.748	-0.983	-0.871	-0.997	-1.003

still decays like N^{-1} . Whereas in column 2, for the eigenvalue approximation, we observe a QMC convergence rate that is slower than N^{-1} , but which is still faster than the expected rate of $-5/6 \approx -0.83$. Surprisingly, for the eigenfunction results we observe a QMC convergence rate that is almost N^{-1} for all of our test cases, regardless of the decays. When $q_a = q'_a = q_b = q'_b = 2$ (last column in Table 2) we would expect summability with $p \approx 1/2$ and convergence arbitrarily close to N^{-1} , which is observed in the errors for both the eigenvalue and eigenfunction approximations. For all of the other cases of $q_a, q'_a, q_b, q'_b \in \{4/3, 2\}$ not presented, some but not all of the decays equal 4/3 and we observe very similar results to column 2 of Table 2. In particular, the convergence rates are almost identical.

6 Conclusion

We have presented a truncation-FE-QMC algorithm for approximating the expectation of the smallest eigenvalue, and linear functionals of the corresponding eigenfunction, of a stochastic eigenvalue problem. Along with the method, we have presented a full analysis of the three components of the error and the final error bound has the same decay rates as the corresponding elliptic source problem for all $p \in (0, 1)$. Throughout the analysis we also proved two key results. First, we proved that the eigenvalue gap is bounded away from 0 uniformly in y . Second, we derived bounds on the derivatives of the smallest eigenvalue and the corresponding eigenfunction.

Our first numerical example presents results that match almost exactly with what is predicted by our theoretical analysis. The second example goes beyond our theoretical setting and illustrates that the method is more robust than our theory suggests.

Regarding future work, one possible avenue is to use higher order QMC rules (which converge at a rate faster than N^{-1}) to approximate the expectation of the minimal eigenvalue. The use of higher order QMC rules requires higher order smoothness of the integrand, but we have already proven bounds on the higher order derivatives of λ_1 (and u_1) in Lemma 3.4. As such, we expect that the theory for using higher order

QMC rules for this eigenvalue problem should work quite easily. Of course, to balance the faster quadrature convergence with the discretisation error one should use a more accurate FE method, which would be more challenging if the domain D was not either convex or had a smooth boundary. Additionally, one could consider embedding our truncation-FE-QMC algorithm in a multilevel framework, which we expect would significantly reduce the cost but would also require further theoretical analysis.

Acknowledgements We thank John Toland (University of Bath) for a very useful discussion which helped us formulate the proof in § 2.2. We gratefully acknowledge the financial support from the Australian Research Council (FT130100655, DP150101770, DP180101356), the Taiwanese National Center for Theoretical Sciences (NCTS)—Mathematics Division, the Statistical and Applied Mathematical Sciences Institute (SAMSI) 2017 Year-long Program on Quasi-Monte Carlo and High-Dimensional Sampling Methods for Applied Mathematics, and the Mathematical Research Institute (MATRIX) 2018 program on the Frontiers of High Dimensional Computation.

A Proof of Theorem 2.6

This proof follows the same structure as [36] to bound the FE error, but in addition we have to track the dependence of all constants on \mathbf{y} . We have opted for the classical min-max argument as opposed to the Babuška–Osborn theory [5–7], because it is more elementary and allows us to determine the explicit influence of the constants. We begin with some preliminary definitions.

For $\mathbf{y} \in U$ and $h > 0$, define the orthogonal projection $P_h : V \rightarrow V_h$ of $u \in V$ with respect to the inner product $\mathcal{A}(\mathbf{y}; \cdot, \cdot)$ on V by

$$\mathcal{A}(\mathbf{y}; u - P_h u, v_h) = 0, \quad \text{for all } v_h \in V_h.$$

Although P_h depends on \mathbf{y} through the bilinear form $\mathcal{A}(\mathbf{y}; \cdot, \cdot)$, we suppress this \mathbf{y} dependence in the notation. Let us denote the energy norm by $\|\cdot\|_{\mathcal{A}(\mathbf{y})} = \sqrt{\mathcal{A}(\mathbf{y}; \cdot, \cdot)}$. Then it is easy to verify that due to the \mathcal{A} -orthogonality of P_h we have

$$\|u - P_h u\|_{\mathcal{A}(\mathbf{y})} = \inf_{v_h \in V_h} \|u - v_h\|_{\mathcal{A}(\mathbf{y})}.$$

Due to (2.8) and (2.9) the energy norm is equivalent to the V -norm:

$$\sqrt{a_{\min}} \|v\|_V \leq \|v\|_{\mathcal{A}(\mathbf{y})} \leq \sqrt{\frac{a_{\min} \bar{\lambda}_1}{\chi_1}} \|v\|_V, \quad \text{for all } v \in V. \quad (\text{A.1})$$

Analogously to the min-max principle (2.13), when the k -dimensional subspaces S_k are restricted to V_h we have the min-max representation for the FE eigenvalues

$$\lambda_{k,h}(\mathbf{y}) = \min_{\substack{S_k \subset V_h \\ \dim(S_k)=k}} \max_{0 \neq u_h \in S_k} \frac{\mathcal{A}(\mathbf{y}; u_h, u_h)}{\mathcal{M}(u_h, u_h)}. \quad (\text{A.2})$$

The strategy of the proof of Theorem 2.6 is to first bound the difference between $u(\mathbf{y})$ and its projection $P_h u(\mathbf{y})$, which is fairly straightforward and follows from the FE results for elliptic source problems. The difficulty lies in the fact that the projections $P_h u(\mathbf{y})$ are not the same as the FE eigenfunctions $u_h(\mathbf{y})$. However, they are close. The next stage of the proof is to bound the eigenvalue and eigenfunction errors (2.33), (2.34) in terms of the projection error. For the eigenvalue error a key ingredient is the classical min-max principle (A.2). Combining the FE error bounds with the projection error bounds yields the required results.

Lemma A.1 *Let $\mathbf{y} \in U$. The projection of $u_1(\mathbf{y}) \in E(\mathbf{y}, \lambda_1(\mathbf{y})) \subset V$ into V_h satisfies*

$$\|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_V \leq Ch, \quad (\text{A.3})$$

where $C > 0$ is independent of \mathbf{y} .

Proof The projection $P_h u_1(\mathbf{y})$ can be equivalently viewed as the FE approximation of an elliptic source problem. Indeed, the variational eigenproblem (2.7) for the eigenpair $(\lambda_1(\mathbf{y}), u_1(\mathbf{y}))$ can be written as

$$\mathcal{A}(\mathbf{y}; u_1(\mathbf{y}), v) = \langle f(\mathbf{y}), v \rangle \quad \text{for all } v \in V,$$

where $f(\mathbf{y}) = \lambda_1(\mathbf{y})c \cdot u_1(\mathbf{y})$ is now assumed fixed. The FE approximation problem of this seeks $\tilde{u}_h(\mathbf{y}) \in V_h$ such that

$$\mathcal{A}(\mathbf{y}; \tilde{u}_h(\mathbf{y}), v_h) = \langle f(\mathbf{y}), v_h \rangle \quad \text{for all } v_h \in V_h,$$

for which due to \mathcal{A} -orthogonality the solution is exactly the projection of the eigenfunction: $\tilde{u}_h(\mathbf{y}) = P_h u_1(\mathbf{y})$. This allows us to bound the projection error using the results from elliptic source problems. In particular, our differential operator fits the setting of affine parametric operator equations from [12]. Since $u_1(\mathbf{y}) \in Z$, it follows that $f(\mathbf{y}) \in L^2(D)$ for all \mathbf{y} . The spaces V_h satisfy the approximation property (2.31), thus by Theorem 2.4 in [12] we have

$$\|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_V \leq C' \|f(\mathbf{y})\|_{L^2(D)} h, \quad (\text{A.4})$$

with constant C' independent of \mathbf{y} and h .

To bound $\|f(\mathbf{y})\|_{L^2(D)}$, we use the upper bound in (2.14), the bound (2.1) on c and then the fact that $\|u_1(\mathbf{y})\|_{\mathcal{M}} = 1$ to give

$$\|f(\mathbf{y})\|_{L^2(D)} \leq \lambda_1(\mathbf{y}) \|c\|_{L^\infty(D)}^{1/2} \|u_1(\mathbf{y})\|_{\mathcal{M}} \leq a_{\max}^{1/2} \overline{\lambda_1}$$

Substituting this into (A.4) we have our desired result with a constant C independent of \mathbf{y} and h . \square

We now estimate the eigenvalue error (Lemma A.2) and the eigenfunction error (Lemma A.4) in terms of the projection error that was estimated in Lemma A.1. Lemma A.3 relates to the gap between the FE eigenvalues and $\lambda_1(\mathbf{y})$, and is used in the proof of Lemma A.4.

Lemma A.2 Let $\mathbf{y} \in U$ and let $h > 0$ be sufficiently small independently of \mathbf{y} . Then

$$|\lambda_1(\mathbf{y}) - \lambda_{1,h}(\mathbf{y})| \leq C \|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_V^2, \quad (\text{A.5})$$

where $C > 0$ is independent of \mathbf{y} .

Proof To prove the result we apply the min–max principle to $\lambda_{1,h}(\mathbf{y})$ and choose the particular subspace $S_{1,h}(\mathbf{y}) := \text{span}(P_h u_1(\mathbf{y}))$, which is a one-dimensional subspace of V provided that $P_h u_1(\mathbf{y}) \neq 0$. To prove that $\dim(S_{1,h}(\mathbf{y})) = 1$ suppose for contradiction that $P_h u_1(\mathbf{y}) = 0$, then by (2.15)

$$1 = \frac{1}{\lambda_1(\mathbf{y})} \|u_1(\mathbf{y})\|_{\mathcal{A}(\mathbf{y})}^2 = \frac{1}{\lambda_1(\mathbf{y})} \|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_{\mathcal{A}(\mathbf{y})}^2.$$

Then using the equivalence of norms (A.1), together with the lower bound in (2.14) and Lemma A.1 we get

$$1 \leq \underbrace{\frac{a_{\min} \overline{\lambda_1}}{\chi_1 \underline{\lambda_1}}}_{=: C'} \|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_V^2 \leq C' C^2 h^2, \quad (\text{A.6})$$

where $C > 0$ is the constant from Lemma A.1 and both C and C' are independent of \mathbf{y} and h . So for $h < (\sqrt{C'} C)^{-1}$, this leads to a contradiction and $\dim(S_{1,h}(\mathbf{y})) = 1$. Therefore, we can choose $S_{1,h}(\mathbf{y})$ in (A.2) to give the inequality

$$\lambda_{1,h}(\mathbf{y}) \leq \max_{0 \neq v_h \in S_{1,h}(\mathbf{y})} \frac{\mathcal{A}(\mathbf{y}; v_h, v_h)}{\mathcal{M}(v_h, v_h)} = \frac{\mathcal{A}(\mathbf{y}; P_h u_1(\mathbf{y}), P_h u_1(\mathbf{y}))}{\mathcal{M}(P_h u_1(\mathbf{y}), P_h u_1(\mathbf{y}))}. \quad (\text{A.7})$$

Using the fact that the norm of the projection is bounded by 1, the numerator is bounded by

$$\mathcal{A}(\mathbf{y}; P_h u_1(\mathbf{y}), P_h u_1(\mathbf{y})) \leq \mathcal{A}(\mathbf{y}; u_1(\mathbf{y}), u_1(\mathbf{y})) = \lambda_1(\mathbf{y}), \quad (\text{A.8})$$

where for the equality in the last step we have used (2.15).

Expanding the denominator in (A.7) gives

$$\begin{aligned} \mathcal{M}(P_h u_1(\mathbf{y}), P_h u_1(\mathbf{y})) &= \|u_1(\mathbf{y})\|_{\mathcal{M}}^2 - 2\mathcal{M}(u_1(\mathbf{y}), u_1(\mathbf{y}) - P_h u_1(\mathbf{y})) \\ &\quad + \|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_{\mathcal{M}}^2. \end{aligned}$$

The first term on the right is 1 by the normalisation of $u_1(\mathbf{y})$ and the last term is positive, so we can bound $\mathcal{M}(P_h u_1(\mathbf{y}), P_h u_1(\mathbf{y}))$ from below by

$$\mathcal{M}(P_h u_1(\mathbf{y}), P_h u_1(\mathbf{y})) \geq 1 - 2\mathcal{M}(u_1(\mathbf{y}), u_1(\mathbf{y}) - P_h u_1(\mathbf{y})).$$

Then, using the fact that $u_1(\mathbf{y})$ is an eigenfunction satisfying (2.7) and using also the \mathcal{A} -orthogonality of the projection P_h we have

$$\begin{aligned}\mathcal{M}(P_h u_1(\mathbf{y}), P_h u_1(\mathbf{y})) &\geq 1 - \frac{2}{\lambda_1(\mathbf{y})} \|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_{\mathcal{A}(\mathbf{y})}^2 \\ &\geq 1 - 2C' \|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_V^2,\end{aligned}\quad (\text{A.9})$$

with $C' > 0$ as in (A.6), using again the lower bound in (2.14) and the equivalence of norms (A.1). For h sufficiently small independently of \mathbf{y} , the right hand side of (A.9) is positive and we can substitute it together with (A.8) into (A.7). Rearranging the resulting inequality gives

$$|\lambda_1(\mathbf{y}) - \lambda_{1,h}(\mathbf{y})| \leq 2C' \lambda_{1,h}(\mathbf{y}) \|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_V^2.$$

Now all that remains is to show that $\lambda_{1,h}(\mathbf{y})$ can be bounded from above independently of \mathbf{y} and h . Analogously to (2.14), using the FE min-max representation (A.2) we have

$$\lambda_{1,h}(\mathbf{y}) \leq \frac{a_{\max}}{a_{\min}} (\chi_{1,h} + 1),$$

where $\chi_{1,h}$ corresponds to the smallest eigenvalue of the negative Laplacian on D , discretised in the FE space V_h with boundary condition (1.2). It is well-known, see e.g. [8, Theorem 10.4], that in the current setting we have $|\chi_{1,h} - \chi_1| \leq C'' h^2$ with $C'' > 0$ independent of h . Thus, for h sufficiently small and independent of \mathbf{y} , there exists a constant such that $\lambda_{1,h}(\mathbf{y})$ can be bounded independent of \mathbf{y} and h as required. \square

Lemma A.3 *Let $\mathbf{y} \in U$ and $h > 0$. Then for all $k = 2, 3, \dots, M_h = \dim(V_h)$*

$$\frac{\lambda_1(\mathbf{y})}{\lambda_{k,h}(\mathbf{y}) - \lambda_1(\mathbf{y})} \leq \rho := \frac{\overline{\lambda_1}}{\delta}, \quad (\text{A.10})$$

with δ as in Proposition 2.4 and $\overline{\lambda_1}$ as in (2.14).

Proof Since the FE eigenvalues converge to the true eigenvalues from above, it follows from Proposition 2.4 and the upper bound on $\lambda_1(\mathbf{y})$ in (2.14) that

$$\lambda_{2,h}(\mathbf{y}) - \lambda_1(\mathbf{y}) \geq \lambda_2(\mathbf{y}) - \lambda_1(\mathbf{y}) \geq \delta \geq \delta \frac{\lambda_1(\mathbf{y})}{\overline{\lambda_1}},$$

which completes the proof because the left hand side of (A.10) attains its maximum when $k = 2$. \square

Lemma A.4 *Let $\mathbf{y} \in U$ and $h > 0$. Then*

$$\|u_1(\mathbf{y}) - u_{1,h}(\mathbf{y})\|_{\mathcal{M}} \leq C \|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_{\mathcal{M}}, \quad (\text{A.11})$$

where C is independent of \mathbf{y} and h .

Proof The FE eigenfunctions form an orthonormal basis for V_h with respect to $\mathcal{M}(\cdot, \cdot)$, and so the projection of $u_1(\mathbf{y})$ can be written as

$$P_h u_1(\mathbf{y}) = \sum_{k=1}^{M_h} \alpha_{k,h}(\mathbf{y}) u_{k,h}(\mathbf{y}),$$

where $\alpha_{k,h}(\mathbf{y}) := \mathcal{M}(P_h u_1(\mathbf{y}), u_{k,h}(\mathbf{y}))$.

The key coefficient in this expansion is $\alpha_{1,h}(\mathbf{y})$. If we assume that $\alpha_{1,h}(\mathbf{y}) \geq 0$ (which we can always ensure by controlling the sign of $u_{1,h}(\mathbf{y})$), then the size of $\alpha_{1,h}(\mathbf{y})$ gives a measure of how close $P_h u_1(\mathbf{y})$ is to $u_{1,h}(\mathbf{y})$. As a first step towards (A.11), consider the difference

$$\begin{aligned} & \|u_1(\mathbf{y}) - \alpha_{1,h}(\mathbf{y}) u_{1,h}(\mathbf{y})\|_{\mathcal{M}} \\ & \leq \|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_{\mathcal{M}} + \|P_h u_1(\mathbf{y}) - \alpha_{1,h}(\mathbf{y}) u_{1,h}(\mathbf{y})\|_{\mathcal{M}}. \end{aligned} \quad (\text{A.12})$$

The first term is exactly our target upper bound in (A.11). The square of the second term can be written as

$$\|P_h u_1(\mathbf{y}) - \alpha_{1,h}(\mathbf{y}) u_{1,h}(\mathbf{y})\|_{\mathcal{M}}^2 = \sum_{k=2}^{M_h} \alpha_{k,h}(\mathbf{y})^2.$$

By [36, Lemma 6.4] (or as is easily verified), we can replace $\alpha_{k,h}(\mathbf{y})$ by

$$\alpha_{k,h}(\mathbf{y}) = \frac{\lambda_1(\mathbf{y})}{\lambda_{k,h}(\mathbf{y}) - \lambda_1(\mathbf{y})} \mathcal{M}(u_1(\mathbf{y}) - P_h u_1(\mathbf{y}), u_{k,h}(\mathbf{y})).$$

Hence, using Lemma A.3 and letting Q_h denote the \mathcal{M} -orthogonal projection onto V_h , we can bound

$$\begin{aligned} \|P_h u_1(\mathbf{y}) - \alpha_{1,h}(\mathbf{y}) u_{1,h}(\mathbf{y})\|_{\mathcal{M}}^2 & \leq \rho^2 \sum_{k=2}^{M_h} \mathcal{M}(u_1(\mathbf{y}) - P_h u_1(\mathbf{y}), u_{k,h}(\mathbf{y}))^2 \\ & = \rho^2 \sum_{k=2}^{M_h} \mathcal{M}(Q_h(u_1(\mathbf{y}) - P_h u_1(\mathbf{y})), u_{k,h}(\mathbf{y}))^2 \\ & \leq \rho^2 \|Q_h(u_1(\mathbf{y}) - P_h u_1(\mathbf{y}))\|_{\mathcal{M}}^2 \\ & \leq \rho^2 \|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_{\mathcal{M}}^2. \end{aligned}$$

Thus, our intermediate bound (A.12) can be written as

$$\|u_1(\mathbf{y}) - \alpha_{1,h}(\mathbf{y}) u_{1,h}(\mathbf{y})\|_{\mathcal{M}} \leq (1 + \rho) \|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_{\mathcal{M}}. \quad (\text{A.13})$$

The final step to prove (A.11) is to show that $\alpha_{1,h}(\mathbf{y})$ is close to 1. To that end, using the reverse triangle inequality and the fact that both $u_1(\mathbf{y})$ and $u_{1,h}(\mathbf{y})$ are normalised we have the following measure of how close $\alpha_{1,h}(\mathbf{y})$ is to 1:

$$\begin{aligned} & \|u_1(\mathbf{y}) - \alpha_{1,h}(\mathbf{y})u_{1,h}(\mathbf{y})\|_{\mathcal{M}} \\ & \geq \left| \|u_1(\mathbf{y})\|_{\mathcal{M}} - \alpha_{1,h}(\mathbf{y})\|u_{1,h}(\mathbf{y})\|_{\mathcal{M}} \right| = |1 - \alpha_{1,h}(\mathbf{y})|. \end{aligned} \quad (\text{A.14})$$

Finally, combining (A.13) and (A.14), it follows by the triangle inequality that

$$\begin{aligned} \|u_1(\mathbf{y}) - u_{1,h}(\mathbf{y})\|_{\mathcal{M}} & \leq \|u_1(\mathbf{y}) - \alpha_{1,h}(\mathbf{y})u_{1,h}(\mathbf{y})\|_{\mathcal{M}} + |1 - \alpha_{1,h}(\mathbf{y})|\|u_{1,h}(\mathbf{y})\|_{\mathcal{M}} \\ & \leq 2(1 + \rho)\|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_{\mathcal{M}}. \end{aligned}$$

□

We now have all of the ingredients needed to prove our FE error bounds.

Proof of Theorem 2.6. The eigenvalue error bound (2.33) follows directly from Lemmas A.1 and A.2. All of the constants involved are independent of \mathbf{y} and h , so the final constant is also.

For the bound on the eigenfunction error, we use [6, Lemma 3.1] to write

$$\|u_1(\mathbf{y}) - u_{1,h}(\mathbf{y})\|_{\mathcal{A}(\mathbf{y})}^2 = \lambda_{1,h}(\mathbf{y}) - \lambda_1(\mathbf{y}) + \lambda_1(\mathbf{y})\|u_1(\mathbf{y}) - u_{1,h}(\mathbf{y})\|_{\mathcal{M}}^2.$$

Then, using the lower bound in the norm equivalence (A.1), as well as the upper bound on $\lambda_1(\mathbf{y})$ in (2.14) and the fact that $\lambda_{1,h} \geq \lambda_1$ this leads to

$$\|u_1(\mathbf{y}) - u_{1,h}(\mathbf{y})\|_V^2 \leq \frac{1}{a_{\min}} |\lambda_{1,h}(\mathbf{y}) - \lambda_1(\mathbf{y})| + \frac{\overline{\lambda_1}}{a_{\min}} \|u_1(\mathbf{y}) - u_{1,h}(\mathbf{y})\|_{\mathcal{M}}^2. \quad (\text{A.15})$$

Now, combining Lemma A.4 with the upper bound in the norm equivalence (2.6) and Poincaré's inequality (2.10) we can estimate

$$\|u_1(\mathbf{y}) - u_{1,h}(\mathbf{y})\|_{\mathcal{M}} \leq C\|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_{\mathcal{M}} \leq C \left(\frac{a_{\max}}{\chi_1} \right)^{1/2} \|u_1(\mathbf{y}) - P_h u_1(\mathbf{y})\|_V,$$

where $C > 0$ is the constant from Lemma A.4. Using this bound in (A.15) together with Lemmas A.1 and A.2 we get

$$\|u_1(\mathbf{y}) - u_{1,h}(\mathbf{y})\|_V \leq C'h,$$

for some constant $C' > 0$ depending only on a_{\max} , a_{\min} and χ_1 , as well as the constants in Lemmas A.1, A.2 and A.4, which are all independent of \mathbf{y} and h . This completes the proof of (2.34).

Having established the error in the V -norm, we use the classical Aubin-Nitsche duality argument to prove the final error bound (2.35). Let $\mathcal{G} \in H^{-1+t}(D)$ and consider the dual problem: Find $v_{\mathcal{G}}(\mathbf{y}) \in V$ such that

$$\mathcal{A}(\mathbf{y}; w, v_{\mathcal{G}}(\mathbf{y})) = \mathcal{G}(w) \quad \text{for all } w \in V. \quad (\text{A.16})$$

Due to the symmetry of $\mathcal{A}(\mathbf{y}; \cdot, \cdot)$, the standard theory for elliptic boundary value problems guarantees the existence of a unique solution $v_{\mathcal{G}}(\mathbf{y}) \in V$ such that $\|v_{\mathcal{G}}(\mathbf{y})\|_V \leq C_1 \|\mathcal{G}\|_{V^*}$. In fact, it can also be shown that $v_{\mathcal{G}}(\mathbf{y}) \in Z^t := V \cap H^{1+t}(D)$ with $\|v_{\mathcal{G}}(\mathbf{y})\|_{Z^t} \leq C_2$. The fact that $C_1, C_2 > 0$ are independent of h is classical; the independence of \mathbf{y} has been shown in [26]. Thus, using the norm equivalences in (A.1) and the best-approximation property of $P_h v_{\mathcal{G}}(\mathbf{y})$ in the energy norm we get

$$\|v_{\mathcal{G}}(\mathbf{y}) - P_h v_{\mathcal{G}}(\mathbf{y})\|_V \leq \sqrt{\frac{\lambda_1}{\chi_1}} \inf_{w_h \in V_h} \|v_{\mathcal{G}}(\mathbf{y}) - w_h\|_V \leq C_3 h^t, \quad (\text{A.17})$$

with $C_3 > 0$ again independent of \mathbf{y} and h , where in the last inequality we have used the approximation property (2.31) and the bound on the Z^t -norm.

Letting $w = u_1(\mathbf{y}) - u_{1,h}(\mathbf{y})$ in (A.16), by the definition of P_h and the boundedness of the bilinear form (2.9) there finally exists a constant $C_4 > 0$ independent of \mathbf{y} and h such that

$$\begin{aligned} |\mathcal{G}(u_1(\mathbf{y})) - \mathcal{G}(u_{1,h}(\mathbf{y}))| &= |\mathcal{G}(u_1(\mathbf{y}) - u_{1,h}(\mathbf{y}))| = |\mathcal{A}(\mathbf{y}; u_1(\mathbf{y}) - u_{1,h}(\mathbf{y}), v_{\mathcal{G}}(\mathbf{y}))| \\ &= |\mathcal{A}(\mathbf{y}; u_1(\mathbf{y}) - u_{1,h}(\mathbf{y}), v_{\mathcal{G}}(\mathbf{y}) - P_h v_{\mathcal{G}}(\mathbf{y}))| \\ &\leq a_{\max}(1 + \chi_1) \|u_1(\mathbf{y}) - u_{1,h}(\mathbf{y})\|_V \|v_{\mathcal{G}}(\mathbf{y}) - P_h v_{\mathcal{G}}(\mathbf{y})\|_V \leq C_4 h^{1+t}. \end{aligned}$$

In the last step, we have used the upper bound (2.34) on the FE error for the eigenfunction together with the dual error bound in (A.17). \square

References

1. Andreev, R., Schwab, Ch.: Sparse tensor approximation of parametric eigenvalue problems. In: Graham, I.G., et al. (eds.) Numerical Analysis of Multiscale Problems. Lecture Notes in Computational Science and Engineering, pp. 203–241. Springer, Berlin (2012)
2. Andrews, B., Clutterbuck, J.: Proof of the fundamental gap conjecture. J. Am. Math. Soc. **24**, 899–916 (2011)
3. Avramova, M.N., Ivanov, K.N.: Verification, validation and uncertainty quantification in multi-physics modeling for nuclear reactor design and safety analysis. Prog. Nucl. Energy **52**(7), 601–614 (2010)
4. Ayres, D.A.F., Eaton, M.D., Hagues, A.W., Williams, M.M.R.: Uncertainty quantification in neutron transport with generalized polynomial chaos using the method of characteristics. Ann. Nucl. Energy **45**, 14–28 (2012)
5. Babuška, I., Osborn, J.: Estimates for the errors in eigenvalue and eigenvector approximation by Galerkin methods, with particular attention to the case of multiple eigenvalues. SIAM J. Numer. Anal. **24**, 1249–1276 (1987)
6. Babuška, I., Osborn, J.: Finite element-Galerkin approximation of eigenvalues and eigenvectors of selfadjoint problems. Math. Comput. **52**, 275–297 (1989)
7. Babuška, I., Osborn, J.: Eigenvalue problems. In: Ciarlet, P.G., Lions, J.L. (eds.) Handbook of Numerical Analysis. Volume 2: Finite Element Methods (Part 1), pp. 641–787. Elsevier Science, Amsterdam (1991)
8. Boffi, D.: Finite element approximation of eigenvalue problems. Acta Numer. **19**, 1–120 (2010)
9. Brezis, H.: Functional Analysis. Sobolev Spaces and Partial Differential Equations. Universitext. Springer, New York (2011)
10. Charrier, J.: Strong and weak error estimates for elliptic partial differential equations with random coefficients. SIAM J. Numer. Anal. **50**, 216–246 (2012)

11. Dick, J., Kuo, F.Y., Sloan, I.H.: High-dimensional integration: the quasi-Monte Carlo way. *Acta Numer.* **22**, 133–288 (2013)
12. Dick, J., Kuo, F.Y., Le Gia, Q.T., Nuyens, D., Schwab, Ch.: Higher order QMC Petrov–Galerkin discretisation for affine parametric operator equations with random field inputs. *SIAM J. Numer. Anal.* **52**, 2676–2702 (2014)
13. Dobson, D.C.: An efficient method for band structure calculations in 2D photonic crystals. *J. Comput. Phys.* **149**(2), 363–376 (1999)
14. Duderstadt, J.J., Hamilton, L.J.: Nuclear Reactor Analysis. Wiley, New York (1976)
15. Ehrlacher, V.: Some Mathematical Models in Quantum Chemistry and Uncertainty Quantification. PhD Thesis, CERMICS, Université Paris-Est (2012)
16. Fumagalli, I., Manzoni, A., Parolini, N., Verani, M.: Reduced basis approximation and a posteriori error estimates for parameterized elliptic eigenvalue problems. *ESAIM: M2AN* **50**, 1857–1885 (2016)
17. Gantner, R.: Dimension truncation in QMC for affine-parametric operator equations. In: Owen, A.B., Glynn, P.W. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2016, pp. 249–264. Springer, Berlin (2018)
18. Ghosh, D., Ghanem, R.G., Red-Horse, J.: Analysis of eigenvalues and modal interaction of stochastic systems. *AIAA J.* **43**(10), 2196–2201 (2005)
19. Giani, S., Graham, I.G.: Adaptive finite element methods for computing band gaps in photonic crystals. *Numer. Math.* **121**(1), 31–64 (2012)
20. Henrot, A.: Extremum Problems for Eigenvalues of Elliptic Operators. Birkhäuser Verlag, Basel (2006)
21. Hörmander, L.: The Analysis of Linear Partial Differential Operators I. Springer, Berlin (2003)
22. Horger, T., Wohlmuth, B., Dickopf, T.: Simultaneous reduced basis approximation of parameterized elliptic eigenvalue problems. *ESAIM: M2AN* **51**, 443–465 (2017)
23. Jamelot, E., Ciarlet Jr., P.: Fast non-overlapping Schwarz domain decomposition methods for solving the neutron diffusion equation. *J. Comput. Phys.* **241**, 445–463 (2013)
24. Kato, T.: Perturbation Theory for Linear Operators. Springer, Berlin (1984)
25. Kuchment, P.: The mathematics of photonic crystals. *SIAM Front. Appl. Math.* **22**, 207–272 (2001)
26. Kuo, F.Y., Schwab, Ch., Sloan, I.H.: Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. *SIAM J. Numer. Anal.* **50**(6), 3351–3374 (2012)
27. Machiels, L., Maday, Y., Oliveira, I.B., Patera, A.T., Rovas, D.V.: Output bounds for reduced-basis approximations of symmetric positive definite eigenvalue problems. *C. R. Acad. Sci. Paris, Sér. I* **331**, 153–158 (2000)
28. Norton, R., Scheichl, R.: Planewave expansion methods for photonic crystal fibres. *Appl. Numer. Math.* **63**, 88–104 (2012)
29. Nuyens, D., Cools, R.: Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces. *Math. Comput.* **75**(254), 903–920 (2006)
30. Nuyens, D., Cools, R.: Fast component-by-component construction of rank-1 lattice rules with a non-prime number of points. *J. Complex.* **22**, 4–28 (2006)
31. Pau, G.S.H.: Reduced-basis method for band structure calculations. *Phys. Rev. E* **79**, 046704 (2007)
32. Pettit, C.L.: Uncertainty quantification in aeroelasticity: recent results and research challenges. *J. Aircr.* **41**(5), 1217–1229 (2004)
33. Scheichl, R.: Parallel Solution of the Transient Multigroup Neutron Diffusion Equations with Multi-Grid and Preconditioned Krylov-Subspace Methods (Master's Thesis). Schriften der Johannes Kepler Universität Linz, Vol. C21, Trauner-Verlag, Linz (1997)
34. Sloan, I.H., Kuo, F.Y., Joe, S.: Constructing randomly shifted lattice rules in weighted Sobolev spaces. *SIAM J. Numer. Anal.* **40**(5), 1650–1665 (2002)
35. Sloan, I.H., Woźniakowski, H.: When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? *J. Complex.* **14**(1), 1–33 (1998)
36. Strang, G., Fix, G.: An Analysis of the Finite Element Method. Wellesley-Cambridge Press, Wellesley, MA (1973)
37. Van den Branden, G.: Nuclear Reactor Theory. Exercises: Part 1(Prof. W. D'haeseleer). Belgian Nuclear Higher Education Network (BNEN) Course (Prof. W. Dhaeseleer), KU Leuven, 2015. available at https://people.mech.kuleuven.be/~william/BNEN/NRT%202014-2015/Exercises%20BNEN%20NRT_WDH_2009_2010.pdf
38. Wachspress, E.L.: Iterative Solution of Elliptic Systems and Applications to the Neutron Diffusion Equations of Reactor Physics. Prentice-Hall Inc, Englewood Cliffs, NJ (1966)

39. Williams, M.M.R.: A method for solving stochastic eigenvalue problems. *Appl. Math. Comput.* **215**(11), 4729–4744 (2010)
40. Williams, M.M.R.: A method for solving stochastic eigenvalue problems II. *Appl. Math. Comput.* **219**(9), 4729–4744 (2013)
41. Zhang, Z., Chen, W., Cheng, X.: Sensitivity analysis and optimization of eigenmode localization in continuum systems. *Struct. Multidiscip. Optim.* **52**(2), 305–317 (2015)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

A. D. Gilbert¹ · I. G. Graham² · F. Y. Kuo³ · R. Scheichl^{1,2} · I. H. Sloan³

I. G. Graham
i.g.graham@bath.ac.uk

F. Y. Kuo
f.kuo@unsw.edu.au

R. Scheichl
r.scheichl@uni-heidelberg.de

I. H. Sloan
i.sloan@unsw.edu.au

¹ Institute for Applied Mathematics and Interdisciplinary Center for Scientific Computing, Universität Heidelberg, 69120 Heidelberg, Germany

² Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK

³ School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2052, Australia