AN ADAPTIVE STOCHASTIC GALERKIN METHOD FOR
RANDOM ELLIPTIC OPERATORS

CLAUDE JEFFREY GITTELSON

Abstract. We derive an adaptive solver for random elliptic boundary value
problems, using techniques from adaptive wavelet methods. Substituting wave-
lets by polynomials of the random parameters leads to a modular solver for the
parameter dependence of the random solution, which combines with any dis-
cretization on the spatial domain. In addition to selecting active polynomial
modes, this solver can adaptively construct a separate spatial discretization
for each of their coefficients. We show convergence of the solver in this general
setting, along with a computable bound for the mean square error, and an
optimality property in the case of a single spatial discretization. Numerical
computations demonstrate convergence of the solver and compare it to a sparse
tensor product construction.

Introduction

Stochastic Galerkin methods have emerged in the past decade as an efficient
solution procedure for boundary value problems depending on random data; see
[14, 32, 2, 60, 23, 15, 31, 28, 6, 5]. These methods approximate the random solution
by a Galerkin projection onto a finite-dimensional space of random fields. This
requires the solution of a single coupled system of deterministic equations for the
coefficients of the Galerkin projection with respect to a predefined set of basis
functions on the parameter domain.

A major remaining obstacle is the construction of suitable spaces in which to
compute approximate solutions. These should be adapted to the stochastic struc-
ture of the equation. Simple tensor product constructions are infeasible due to the
high dimensionality of the parameter domain in the case of input random fields
with low regularity.

Parallel to but independently from the development of stochastic Galerkin meth-
ods, a new class of adaptive methods has emerged, which are set not in the con-
tinuous framework of a boundary value problem, but rather on the level of coeffi-
cients with respect to a hierarchic Riesz basis, such as a wavelet basis. Due to the
norm equivalences constitutive of Riesz bases, errors and residuals in appropriate
sequence spaces are equivalent to those in physically meaningful function spaces.
This permits adaptive wavelet methods to be applied directly to a large class of
equations, provided that a suitable Riesz basis is available.
For symmetric elliptic problems, the error of the Galerkin projection onto the
span of a set of coefficients can be estimated using a sufficiently accurate approxi-
mentation of the residual of a previously computed approximate solution; see [8, 19, 16].
This results in a sequence of finite-dimensional linear equations with successively
larger sets of active coefficients.

We use techniques from these adaptive wavelet methods to derive an adaptive
solver for random symmetric elliptic boundary value problems. In place of wavelets,
we use an orthonormal polynomial basis on the parameter domain. The coefficients
of the random solution with respect to this basis are deterministic functions on the
spatial domain.

Adaptive wavelet methods extend to this vector setting, and lead to a modular
solver which can be coupled with any discretization of or solver for the deterministic
problem. We consider adaptive finite elements with a residual-based a posteriori
error estimator.

We review random operator equations in Section 1. In particular, we derive
the weak formulation of such equations, construct orthonormal polynomials on the
parameter domain, and recast the weak formulation as a bi-infinite operator matrix
equation for the coefficients of the random solution with respect to this polynomial
basis. We refer to [22] for further details.

A crucial ingredient in adaptive wavelet methods is the approximation of the
residual. We study this for the setting of stochastic operator equations in Section 2.
The resulting adaptive solver is presented in Section 3. We show convergence of the
method, and provide a reliable error bound. Optimality properties are discussed in
Section 4 for the special case of a fixed spatial discretization.

Finally, in Section 5, we apply the method to a simple elliptic equation. We
discuss a suitable a posteriori finite element error estimator, and present numeri-
cal computations. These demonstrate the convergence of our solver and compare
the adaptively constructed discretizations with the a priori adapted sparse tensor
product construction from [5]; we refer to [21] for a comparison with other adaptive
solvers. We discuss the empirical convergence behavior in the light of the theoretical
approximation results in [11, 10].

1. Stochastic operator equations

1.1. Pointwise definition. Let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and let $V$ be a separable Hilbert space
over $\mathbb{K}$. We denote by $V^*$ the space of all continuous antilinear functionals on $V$. Furthermore, $\mathcal{L}(V, V^*)$ is the Banach space of bounded linear maps from $V$ to $V^*$.

We consider operator equations depending on a parameter in $\Gamma := [-1, 1]^\infty$.
Given

\begin{equation}
A: \Gamma \to \mathcal{L}(V, V^*) \quad \text{and} \quad f: \Gamma \to V^*,
\end{equation}

we wish to determine

\begin{equation}
u: \Gamma \to V, \quad A(y)u(y) = f(y) \quad \forall y \in \Gamma.
\end{equation}

Let $D \in \mathcal{L}(V, V^*)$ be the Riesz isomorphism, i.e., $\langle D \cdot, \cdot \rangle$ is the scalar product in $V$. We decompose $A$ as

\begin{equation}
A(y) = D + R(y) \quad \forall y \in \Gamma
\end{equation}
and assume that $R(y)$ is linear in $y \in \Gamma$,

$$
R(y) = \sum_{m=1}^{\infty} y_m R_m \quad \forall y = (y_m)_{m=1}^{\infty} \in \Gamma;
$$

e.g., as in [5, 6, 11, 10, 28]. Here, each $R_m$ is in $\mathcal{L}(V, V^*)$. We assume $(R_m)_m \in \ell^1(\mathbb{N}; \mathcal{L}(V, V^*))$, and there is a $\gamma \in [0, 1]$ such that $\|R(y)\|_{V \rightarrow V^*} \leq \gamma$ for all $y \in \Gamma$. By [22, Proposition 1.2], this ensures existence and uniqueness of the solution of (1.1). For simplicity, we also assume that the sequence $(\|R_m\|_{V \rightarrow V^*})_{m=1}^{\infty}$ is nonincreasing.

1.2. Weak formulation. Let $\pi$ be a probability measure on the parameter domain $\Gamma$ with Borel $\sigma$-algebra $\mathcal{B}(\Gamma)$. We assume that the map $\Gamma \ni y \mapsto A(y) v(y)$ is measurable for any measurable $v : \Gamma \rightarrow V$. Then

$$
A : \mathcal{L}^2(\Gamma; V) \rightarrow \mathcal{L}^2(\Gamma; V^*) , \quad v \mapsto [y \mapsto A(y) v(y)] ,
$$
is well defined and continuous. We assume also that $f \in \mathcal{L}^2(\Gamma; V^*)$.

The weak formulation of (1.2) is to find $u \in \mathcal{L}^2(\Gamma; V)$ such that

$$
\int_{\Gamma} \langle A(y)u(y), v(y) \rangle \, d\pi(y) = \int_{\Gamma} \langle f(y), v(y) \rangle \, d\pi(y) \quad \forall v \in \mathcal{L}^2(\Gamma; V) .
$$
The left term in (1.6) is the duality pairing in $\mathcal{L}^2(\Gamma; V)$ of $Au$ with the test function $v$, and the right term is the duality pairing of $f$ with $v$. We follow the convention that the duality pairing is linear in the first argument and antilinear in the second.

By [22, Theorem 1.4], the solution $u$ of (1.2) is in $\mathcal{L}^2(\Gamma; V)$, and it is the unique solution of (1.6). In particular, the operator $A$ is boundedly invertible.

We define the multiplication operators

$$
K_m : \mathcal{L}^2(\Gamma) \rightarrow \mathcal{L}^2(\Gamma) , \quad v(y) \mapsto y_m v(y) , \quad m \in \mathbb{N} .
$$
Since $y_m$ is real and $|y_m|$ is less than one, $K_m$ is symmetric and has norm at most one.

By separability of $V$, the Lebesgue–Bochner space $\mathcal{L}^2(\Gamma; V)$ is isometrically isomorphic to the Hilbert tensor product $\mathcal{L}^2(\Gamma) \otimes V$, and similarly for $V^*$ in place of $V$. Using these identifications, we expand $A$ as $A = D + R$ with

$$
D := \text{id}_{\mathcal{L}^2(\Gamma)} \otimes D \quad \text{and} \quad R := \sum_{m=1}^{\infty} K_m \otimes R_m .
$$
This sum converges in $\mathcal{L}(\mathcal{L}^2(\Gamma; V), \mathcal{L}^2(\Gamma; V^*))$ due to the assumption that $(R_m)_m \in \ell^1(\mathbb{N}; \mathcal{L}(V, V^*))$.

Lemma 1.1. $\|R\|_{\mathcal{L}^2(\Gamma; V) \rightarrow \mathcal{L}^2(\Gamma; V^*)} \leq \gamma < 1$.

Proof. We note that, as in (1.5), $\langle Rv \rangle(y) = R(y) v(y)$ for all $v \in \mathcal{L}^2(\Gamma; V)$ and $y \in \Gamma$. Therefore, using the assumption $\|R(y)\|_{V \rightarrow V^*} \leq \gamma$,

$$
\|Rv\|_{\mathcal{L}^2(\Gamma; V^*)}^2 = \int_{\Gamma} \|R(y) v(y)\|_{V^*}^2 \, d\pi(y) \leq \int_{\Gamma} \|R(y)\|_{V \rightarrow V^*}^2 \|v(y)\|_V^2 \, d\pi(y) ,
$$
and the assertion follows using the assumption $\|R(y)\|_{V \rightarrow V^*} \leq \gamma$. \qed
1.3. **Orthonormal polynomial basis.** In order to construct an orthonormal polynomial basis of $L^2_\pi(\Gamma)$, we assume that $\pi$ is a product measure. Let

$$\pi = \bigotimes_{m=1}^{\infty} \pi_m$$

for probability measures $\pi_m$ on $([-1, 1], \mathcal{B}([-1, 1]))$; see e.g. [4, Section 9] for a general construction of arbitrary products of probability measures. We assume that the support of $\pi_m$ in $[-1, 1]$ has infinite cardinality.

For all $m \in \mathbb{N}$, let $(P^m_n)_{n=0}^{\infty}$ be an orthonormal polynomial basis of $L^2_{\pi_m}([-1, 1])$, with $\deg P^m_n = n$. Such a basis is given by the three term recursion

$$P^m_{-1} := 0, \quad P^m_0 := 1$$

and

$$\beta^m_n P^m_n(\xi) := (\xi - \alpha^m_{n-1})P^m_{n-1}(\xi) - \beta^m_{n-1} P^m_{n-2}(\xi), \quad n \in \mathbb{N},$$

with

$$\alpha^m_n := \int_{-1}^{1} \xi P^m_n(\xi)^2 \, d\pi_m(\xi) \quad \text{and} \quad \beta^m_n := \frac{c^m_{n-1}}{c^m_n},$$

where $c^m_n$ is the leading coefficient of $P^m_n$, $\beta^m_0 := 1$, and $P^m_0$ is chosen as normalized in $L^2_{\pi_m}(0, 1)$ with a positive leading coefficient.

We define the set of finitely supported sequences in $\mathbb{N}_0$ as

$$\Lambda := \{\nu \in \mathbb{N}_0^\mathbb{N}; \# \text{ supp } \nu < \infty\},$$

where the support is defined by

$$\text{ supp } \nu := \{m \in \mathbb{N}; \nu_m \neq 0\}, \quad \nu \in \mathbb{N}_0^\mathbb{N}.$$

Then countably infinite tensor product polynomials are given by

$$P := (P_\nu)_{\nu \in \Lambda}, \quad P_\nu := \bigotimes_{m=1}^{\infty} P^m_{\nu_m}, \quad \nu \in \Lambda.$$

Note that each of these functions depends on only finitely many dimensions,

$$P_\nu(y) = \prod_{m=1}^{\infty} P^m_{\nu_m}(y_m) = \prod_{m \in \text{ supp } \nu} P^m_{\nu_m}(y_m), \quad \nu \in \Lambda,$$

since $P^m_0 = 1$ for all $m \in \mathbb{N}$.

For example, by [22, Theorem 2.8], $P$ is an orthonormal basis of $L^2_\pi(\Gamma)$. By Parseval’s identity, this is equivalent to the statement that the map

$$T: \ell^2(\Lambda) \rightarrow L^2_\pi(\Gamma), \quad (c_\nu)_{\nu \in \Lambda} \mapsto \sum_{\nu \in \Lambda} c_\nu P_\nu,$$

is a unitary isomorphism. The inverse of $T$ is

$$T^{-1} = T^*: L^2_\pi(\Gamma) \rightarrow \ell^2(\Lambda), \quad g \mapsto \left(\int_{\Gamma} g(y)P_\nu(y) \, d\pi(y)\right)_{\nu \in \Lambda}.$$
1.4. Bi-infinite operator matrix equation. We use the isomorphism $T$ from (1.16) to recast the weak stochastic operator equation (1.6) as an equivalent discrete operator equation. Since $T$ is a unitary map from $\ell^2(\Lambda)$ to $L^2_{\pi}(\Gamma)$, the tensor product operator $T_V := T \otimes \text{id}_V$ is an isometric isomorphism from $\ell^2(\lambda; V)$ to $L^2_{\pi}(\Gamma; V)$. By definition, $w \in L^2_{\pi}(\Gamma; V)$ and $w = (w_\nu)_{\nu \in \Lambda} \in \ell^2(\Lambda; V)$ are related by $w = T_V w$ if

$$w(y) = \sum_{\nu \in \Lambda} w_\nu P_\nu(y) \quad \text{or} \quad w_\nu = \int_\Gamma w(y) P_\nu(y) \, d\pi(y) \quad \forall \nu \in \Lambda,$$

and either of these properties implies the other. The series in (1.18) converges unconditionally in $L^2_{\pi}(\Gamma; V)$, and the integral can be interpreted as a Bochner integral in $V$.

Let $A := T_V^* A T_V$ and $f := T_V^* f$. Then $u = T_V u$ for $u \in \ell^2(\Lambda; V)$ with

$$Au = f$$

since $u \in L^2_{\pi}(\Gamma; V)$ satisfies $Au = f$.

By definition, $A$ is a boundedly invertible linear map from $\ell^2(\Lambda)$ to $\ell^2(\Lambda; V^*)$. It can be interpreted as a bi-infinite operator matrix

$$A = [A_{\nu \mu}]_{\nu, \mu \in \Lambda}, \quad A_{\nu \mu} : V \to V^*,$$

with entries

$$A_{\nu \nu} = D + \sum_{m=1}^{\infty} \alpha_{\nu m} R_m, \quad \nu \in \Lambda,$$

$$A_{\nu \mu} = \beta_{\max(\nu m, \mu m)} R_m, \quad \nu, \mu \in \Lambda, \quad \nu - \mu = \pm \epsilon_m,$$

and $A_{\nu \mu} = 0$ otherwise, where $\epsilon_m$ denotes the Kronecker sequence with $(\epsilon_m)_n = \delta_{mn}$. If $\pi_m$ is a symmetric measure on $[-1, 1]$ for all $m \in \mathbb{N}$, then $\alpha_{\nu m} = 0$ for all $m$ and $n$, and thus $A_{\nu \nu} = D$. We refer to [22, 20] for details.

Similarly, the operator $R := T_V^* R T_V$ can be interpreted as a bi-infinite operator matrix $R = [R_{\mu \nu}]$ with $R_{\mu \nu} = A_{\nu \mu}^* - D$ and $R_{\mu \nu} = A_{\nu \mu}$ for $\nu \neq \mu$.

Let $K_m = T^* K_m T \in \mathcal{L}(\ell^2(\Lambda))$. Due to the three term recursion (1.10),

$$K_m c(\mu) = \beta_{\mu m + 1} c_{\mu + 1} + \alpha_{\mu m} c_{\mu} + \beta_{\mu m} c_{\mu - 1}, \quad \mu \in \Lambda,$$

for $c = (c_{\mu})_{\mu \in \Lambda} \in \ell^2(\Lambda)$, where $c_{\mu} := 0$ if $\mu_{\mu} < 0$ for any $m \in \mathbb{N}$. Furthermore, $K^*_m = K_m$ and $\|K_m\|_{\ell^2(\Lambda) \to \ell^2(\Lambda)} \leq 1$.

Using the maps $K_m$, $R$ can be written succinctly as

$$R = \sum_{m=1}^{\infty} K_m \otimes R_m,$$

with unconditional convergence in $\mathcal{L}(\ell^2(\Lambda; V) , \ell^2(\Lambda; V^*)$. By Lemma (1.1)

$$\|R\|_{\ell^2(\Lambda; V) \to \ell^2(\Lambda; V^*)} \leq \gamma < 1.$$

In particular, $\|A\| \leq (1 + \gamma)$ and $\|A^{-1}\| \leq (1 - \gamma)^{-1}$.

We also define the operator $D := T_V^* D T_V$. This is just the Riesz isomorphism from $\ell^2(\Lambda; V)$ to $\ell^2(\Lambda; V^*)$. By [22, Proposition 2.10],

$$(1 - \gamma) D \leq A \leq (1 + \gamma) D \quad \text{and} \quad \frac{1}{1 + \gamma} D^{-1} \leq A^{-1} \leq \frac{1}{1 - \gamma} D^{-1}.$$
In particular, using \( A = A A^{-1} A \), we have
\[
\frac{1}{1 + \gamma} A D^{-1} A \leq A \leq \frac{1}{1 - \gamma} A D^{-1} A .
\]

1.5. **Galerkin projection.** Let \( W \) be a closed subspace of \( L^2_{\pi}(\Gamma; V) \). The Galerkin solution \( \bar{u} \in W \) is defined through the linear variational problem
\[
\int_{\Gamma} \langle A(y) \bar{u}(y), w(y) \rangle \; d\pi(y) = \int_{\Gamma} \langle f(y), w(y) \rangle \; d\pi(y) \quad \forall w \in W .
\]
Existence, uniqueness and quasi-optimality of \( \bar{u} \) follow since \( A \) induces an inner product on \( L^2_{\pi}(\Gamma; V) \) that is equivalent to the standard inner product; see [22, Proposition 1.5].

For all \( \nu \in \Lambda \), let \( W_{\nu} \) be a finite dimensional subspace of \( V \), such that \( W_{\nu} \neq \{0\} \) for only finitely many \( \nu \in \Lambda \). It is particularly useful to consider spaces \( W \) of the form
\[
W := \sum_{\nu \in \Lambda} W_{\nu} P_{\nu} .
\]
The Galerkin operator on such a space has a similar structure to (1.20), with \( A_{\nu\mu} \) replaced by its representation on suitable subspaces \( W_{\nu} \) of \( V \); see [22, Section 2].

2. **Approximation of the residual**

2.1. **Adaptive application of the stochastic operator.** We construct a sequence of approximations of \( R \) by truncating the series (1.23). For all \( M \in \mathbb{N} \), let
\[
R_{[M]} := \sum_{m=1}^{M} K_{m} \otimes R_{m} ,
\]
and \( R_{[0]} := 0 \). For all \( M \in \mathbb{N} \), let \( \bar{e}_{R_{M}} \) be given such that
\[
\| R - R_{[M]} \|_{\ell^2(A; V) \rightarrow \ell^2(A; V^*)} \leq \bar{e}_{R_{M}} .
\]
For example, these bounds can be chosen as
\[
\bar{e}_{R_{M}} := \sum_{m=M+1}^{\infty} \| R_{m} \|_{V \rightarrow V^*} .
\]
We assume that \( (\bar{e}_{R_{M}})_{M=0}^{\infty} \) is nonincreasing and converges to 0, and also that the sequence of differences \( (\bar{e}_{R_{M}} - \bar{e}_{R_{M+1}})_{M=0}^{\infty} \) is nonincreasing.

We consider a partitioning of a vector \( w \in \ell^2(\Lambda) \) into \( w_{[p]} := w|_{A_{p}}, p = 1, \ldots, P \), for disjoint index sets \( A_{p} \subset \Lambda \). This can be approximate in that \( w_{[1]} + \cdots + w_{[P]} \) only approximates \( w \) in \( \ell^2(\Lambda) \). We think of \( w_{[1]} \) as containing the largest elements of \( w, w_{[2]} \) the next largest, and so on.

Such a partitioning can be constructed by the approximate sorting algorithm
\[
\text{BucketSort}[w, \epsilon] \mapsto \{ (w_{[p]})_{P=1}^{P}, (A_{p})_{P=1}^{P} \} ,
\]
which, given a finitely supported \( w \in \ell^2(\Lambda) \) and a threshold \( \epsilon > 0 \), returns index sets
\[
A_{p} := \left\{ \mu \in \Lambda ; |w_{\mu}| \in \left( 2^{-p/2} \| w \|_{\ell^\infty}, 2^{-(p-1)/2} \| w \|_{\ell^\infty} \right) \right\} .
\]
and \( w[p] := w|_{\Lambda_p} \); see \([24, 3, 19, 16]\). The integer \( P \) is minimal with
\[
2^{-P/2} \|w\|_{\ell^\infty(A)} \sqrt{\# \supp w} \leq \epsilon.
\]
By \([19, \text{Rem. 2.3}]\) or \([16, \text{Prop. 4.4}]\), the number of operations and storage locations required by a call of \( \text{BucketSort}[w, \epsilon] \) is bounded by
\[
\# \supp w + \max(1, \lceil \log(\|w\|_{\ell^\infty(A)} \sqrt{\# \supp w/\epsilon}) \rceil).
\]
This analysis uses that every \( w_\mu, \mu \in \Lambda \), can be mapped to \( p \) with \( \mu \in \Lambda_p \) in constant time by evaluating
\[
p := \left\lfloor 1 + 2 \log_2 \left( \frac{\|w\|_{\ell^\infty(A)}}{|w_\mu|} \right) \right\rfloor.
\]
Alternatively, any standard comparison-based sorting algorithm can be used to construct the partitioning of \( w \), albeit with an additional logarithmic factor in the complexity.

\[
\text{Apply}_{R}[v, \epsilon] \mapsto z
\]
\[
[\cdot,(\Lambda_p)_p^P] \leftarrow \text{BucketSort} \left[ (\|v_\mu\|_V)_{\mu \in \Lambda}, \frac{\epsilon}{2\bar{e}_{R,0}} \right]
\]
for \( p = 1, \ldots, P \) do \( v[p] \leftarrow (v_\mu)_{\mu \in \Lambda_p} \)
Compute the minimal \( \ell \in \{0,1,\ldots,P\} \) s.t. \( \bar{e}_{R,0} \left\| v - \sum_{p=1}^\ell v[p] \right\|_{\ell^2(\Lambda;V)} \leq \frac{\epsilon}{2} \)
for \( p = 1, \ldots, \ell \) do \( M_p \leftarrow 0 \)
while \( \sum_{p=1}^\ell \bar{e}_{R,M_p} \left\| v[p] \right\|_{\ell^2(\Lambda;V)} > \epsilon - \delta \) do
\[
q \leftarrow \arg\max_{p=1,\ldots,\ell} (\bar{e}_{R,M_p} - \bar{e}_{R,M_p+1}) \left\| v[p] \right\|_{\ell^2(\Lambda;V)} / \# \Lambda_p
\]
\[
M_q \leftarrow M_q + 1
\]
\[
z = (z_\nu)_{\nu \in \Lambda} \leftarrow 0
\]
for \( p = 1, \ldots, \ell \) do
\[
\text{forall the } \mu \in \Lambda_p \text{ do}
\]
\[
\text{for } m = 1, \ldots, M_p \text{ do}
\]
\[
 w \leftarrow R_{m} v_{\mu}
\]
\[
z_{\mu+\epsilon_m} \leftarrow z_{\mu+\epsilon_m} + \beta_{\mu+\epsilon_m} w
\]
if \( \mu_m \geq 1 \) then \( z_{\mu-\epsilon_m} \leftarrow z_{\mu-\epsilon_m} + \beta_{\mu-\epsilon_m} w
\]
if \( \alpha_{\mu_m} \neq 0 \) then \( z_{\mu} \leftarrow z_{\mu} + \alpha_{\mu_m} w
\]

The routine \( \text{Apply}_{R}[v, \epsilon] \) adaptively approximates \( Rv \) in three distinct steps. First, the elements of \( v \) are grouped according to their norm. Elements smaller than a certain tolerance are discarded. This truncation of the vector \( v \) produces an error of at most \( \delta \leq \epsilon/2 \).

Next, a greedy algorithm is used to assign to each segment \( v[p] \) of \( v \) an approximation \( R_{[M_p]} \) of \( R \). Starting with \( R_{[M_p]} = 0 \) for all \( p = 1, \ldots, \ell \), these approximations are refined iteratively until an estimate of the error is smaller than \( \epsilon - \delta \).
Finally, the operations determined by the previous two steps are performed. Each multiplication $R_m v_\mu$ is performed just once, and copied to the appropriate entries of $z$.

**Proposition 2.1.** For any finitely supported $v \in \ell^2(\Lambda; V)$ and any $\epsilon > 0$, $\text{Apply}_R[v, \epsilon]$ produces a finitely supported $z \in \ell^2(\Lambda; V^*)$ with

$$
\# \text{supp } z \leq 3 \sum_{p=1}^\ell M_p \# \Lambda_p
$$

and

$$
\| Rv - z \|_{\ell^2(\Lambda; V^*)} \leq \delta + \eta_M \leq \epsilon , \quad \eta_M := \sum_{p=1}^\ell \bar{e}_{R,M_p} \| v[p] \|_{\ell^2(\Lambda,V)} ,
$$

where $M_p$ refers to the final value of this variable in the call of $\text{Apply}_R$. The total number of products $R_m v_\mu$ computed in $\text{Apply}_R[v, \epsilon]$ is $\sigma_M := \sum_{p=1}^\ell M_p \# \Lambda_p$.

Furthermore, the vector $M = (M_p)_{p=1}^\ell$ is optimal in the sense that if $N = (N_p)_{p=1}^\ell$ with $\sigma_N \leq \sigma_M$, then $\eta_N \geq \eta_M$, and if $\sigma_N \leq \sigma_M$, then $\eta_N \geq \eta_M$.

**Proof.** The estimate (2.9) follows from the fact that each $K_m$ has at most three nonzero entries per column; see (1.22). Since $\| R \|_{\ell^2(\Lambda; V) \rightarrow \ell^2(\Lambda; V^*)} \leq \bar{e}_{R,0}$,

$$
\left\| Rv - R \sum_{p=1}^\ell v[p] \right\|_{\ell^2(\Lambda; V^*)} \leq \bar{e}_{R,0} \left\| v - \sum_{p=1}^\ell v[p] \right\|_{\ell^2(\Lambda; V)} = \delta \leq \epsilon / 2 .
$$

Due to (2.2) and the termination criterion in the greedy subroutine of $\text{Apply}_R$,

$$
\sum_{p=1}^\ell \left\| Rv[p] - R[M_p]v[p] \right\|_{\ell^2(\Lambda; V^*)} \leq \sum_{p=1}^\ell \bar{e}_{R,M_p} \| v[p] \|_{\ell^2(\Lambda; V)} \leq \epsilon - \delta .
$$

For the optimality property of the greedy algorithm, we refer to the more general statement [20, Theorem 4.1.5].

### 2.2. Computation of the residual.

We assume a solver for $D$ is available such that for any $g \in V^*$ and any $\epsilon > 0$,

$$
\text{Solve}_D[g, \epsilon] \mapsto v , \quad \| v - D^{-1}g \|_V \leq \epsilon .
$$

For example, $\text{Solve}_D$ could be an adaptive wavelet method (see e.g. [8, 9, 19]), an adaptive frame method (see e.g. [27, 12, 13]), or a finite element method with a posteriori error estimation; see e.g. [17, 25, 7].

Furthermore, we assume that a routine

$$
\text{RHS}_f[\epsilon] \mapsto \hat{f}
$$

is available to compute approximations $\hat{f} = (\hat{f}_\nu)_{\nu \in \Lambda}$ of $f$ with $\# \text{supp } \hat{f} < \infty$ and

$$
\| f - \hat{f} \|_{\ell^2(\Lambda; V^*)} \leq \epsilon
$$

for any $\epsilon > 0$.

The routine $\text{Residual}_{A,f}$ approximates the residual $f - Av$ up to a prescribed relative tolerance.
\begin{verbatim}
Residual_{A, f}[\epsilon, v, \eta_0, \chi, \omega, \alpha, \beta] \mapsto [w, \eta, \zeta]
\end{verbatim}

\textbf{Proposition 2.2.} For any finitely supported \(v = (v_\nu)_{\nu \in \Lambda} \in \ell^2(\Lambda; V), \epsilon > 0, \eta_0 \geq 0, \chi > 0, \omega > 0, 0 < \alpha < 1 \) and \(0 < \beta < 1\), a call of \textbf{Residual}_{A, f}[\epsilon, v, \eta_0, \chi, \omega, \alpha, \beta] \text{computes} \(w \in \ell^2(\Lambda; V), \eta \geq 0\) and \(\zeta \geq 0\) with
\begin{equation}
\|g - (f - Rv)\|_{\ell^2(\Lambda; V^*)} \leq (1 - \alpha)\zeta.
\end{equation}
Furthermore, using \(\|w - D^{-1}g\|_{\ell^2(\Lambda; V)} \leq \alpha\zeta\),
\begin{equation}
\|w - D^{-1}(f - Rv)\|_{\ell^2(\Lambda; V)} \leq \|w - D^{-1}g\|_{\ell^2(\Lambda; V)} + \|g - (f - Rv)\|_{\ell^2(\Lambda; V^*)} \leq \zeta.
\end{equation}
The rest of (2.14) follows by triangle inequality with
\[\|r\|_{\ell^2(\Lambda; V^*)} = \|D^{-1}r\|_{\ell^2(\Lambda; V)}.\]

\textbf{Remark 2.3.} The tolerance \(\zeta\) in \textbf{Residual}_{A, f} is initialized as the product of an initial estimate \(\eta_0\) of the residual and a parameter \(\chi\). The update
\begin{equation}
\zeta \leftarrow \omega \frac{1 - \omega}{1 + \omega} (\eta + \zeta) =: \zeta_1
\end{equation}
ensures a geometric decrease of \(\zeta\) since if \(\zeta > \omega \eta\), then
\begin{equation}
\zeta_1 = \omega \frac{1 - \omega}{1 + \omega} (\eta + \zeta) < \frac{1 - \omega}{1 + \omega} (\zeta + \omega \zeta) = (1 - \omega)\zeta.
\end{equation}
Therefore, the total computational cost of the routine is proportional to that of the final iteration of the loop. Furthermore, if \(\zeta > \omega \eta\), then also
\begin{equation}
\zeta_1 = \omega \frac{1 - \omega}{1 + \omega} (\eta + \zeta) > \omega (1 - \omega) \eta > \omega (\eta - \zeta).
\end{equation}
The term \(\eta - \zeta\) in the last expression of (2.17) is a lower bound for the true residual \(\|r\|_{\ell^2(\Lambda; V^*)}\). In this sense, the prescription (2.15) does not select an unnecessarily small tolerance.
Finally, if \(\zeta \leq 2\omega (1 - \omega)^{-1} \eta\), then \(\zeta_1 \leq \omega \eta\). If the next value of \(\eta\) is greater than or equal to the current value, this ensures that the termination criterion is met in the next iteration. For example, under the mild condition \(\zeta \leq (1 + 4 \omega - \omega^2) (1 - \omega)^{-2} \eta\), we have \(\zeta_1 \leq 2\omega (1 - \omega)^{-1} \eta\). The loop can therefore be expected to terminate within three iterations.
Remark 2.4. In Residual\textsubscript{A.f}, the tolerances of Solve\textsubscript{D} are chosen such that the error tolerance $\alpha \zeta$ is equidistributed among all the nonzero indices of $w$. This property is not required anywhere; Proposition 2.2 only uses that the total error in the computation of $D^{-1}g$ is no more than $\alpha \zeta$. Indeed, other strategies for selecting tolerances, e.g., based on additional a priori information, may be more efficient. Equidistributing the error among all the indices is a simple, practical starting point.

3. An adaptive solver

3.1. Refinement strategy. We use the approximation of the residual described in Section 2 to refine a Galerkin subspace $W \subset L_2^2(\Gamma; V)$ of the form (1.28). For some approximate solution $v$ with $T \nu v \in W$, let $w$ be the approximation of $D^{-1}(f - Rv)$ computed by Residual\textsubscript{A.f}. We construct a space

$$W := \sum_{\mu \in A} \bar{W}_\mu P_\mu \supset W,$$

with $\bar{W}_\mu \subset V$ finite-dimensional, such that $w$ can be approximated sufficiently in $\bar{W}$. A simple choice is $\bar{W}_\mu := W_\mu + \text{span} w_\mu$, where $W = \sum_{\mu \in A} W_\mu P_\mu$.

We consider a multilevel setting. For each $\mu \in \text{supp} w \subset A$, let $W_\mu =: W_\mu^0 \subset W_\mu^1 \subset \cdots$ be a scale of finite-dimensional subspaces of $V$ such that $\bigcup_{i=0}^\infty W_\mu^i$ is dense in $V$. To each space, we associate a cost $\dim W_\mu^i$ and an error $\|w_\mu - \Pi_\mu^i w_\mu\|_V$, where $\Pi_\mu^i$ denotes the orthogonal projection in $V$ onto $W_\mu^i$. In the construction of $\bar{W}$, we use a greedy algorithm to minimize the dimension of $\bar{W}$ under a constraint on the approximation error of $w$.

Proposition 3.1. If for every $\mu \in \text{supp} w$,

$$\frac{\|\Pi_\mu^{i+1} w_\mu - \Pi_\mu^i w_\mu\|_V^2}{\dim(W_\mu^{i+1} \setminus W_\mu^i)} \geq \frac{\|\Pi_\mu^{j+1} w_\mu - \Pi_\mu^j w_\mu\|_V^2}{\dim(W_\mu^{j+1} \setminus W_\mu^j)} \quad \forall i \leq j,$$

where $\Pi_\mu^i$ denotes the orthogonal projection in $V$ onto $W_\mu^i$. In the construction of $\bar{W}$, we use a greedy algorithm to minimize the dimension of $\bar{W}$ under a constraint on the approximation error of $w$. 

\[
\text{Refine}_D(W, w, \varepsilon) \mapsto (W, \bar{W}, \bar{w}, \varrho)
\]

forall the $\mu \in \text{supp} w$ do $j_\mu \leftarrow 0$

while $\sum_{\mu \in \text{supp} w} \|w_\mu - \Pi_\mu w_\mu\|_V^2 > \varepsilon^2$ do

$\nu \leftarrow \arg\max_{\mu \in \text{supp} w} \|\Pi_\mu^{j_\mu+1} w_\mu - \Pi_\mu^{j_\mu} w_\mu\|_V^2 / \dim(W_\mu^{j_\mu+1} \setminus W_\mu^{j_\mu})$

$j_\nu \leftarrow j_\nu + 1$

forall the $\mu \in \text{supp} w$ do

$W_\mu \leftarrow W_\mu^{j_\mu}$

$\bar{w}_\mu \leftarrow \Pi_\mu^{j_\mu} w_\mu$

$\varrho \leftarrow \left(\sum_{\mu \in \text{supp} w} \|w_\mu - \bar{w}_\mu\|_V^2\right)^{1/2}$

Proposition 3.1. If for every $\mu \in \text{supp} w$,
then for any $\epsilon \geq 0$, a call of $\text{Refine}_D[\mathcal{W}, \mathbf{w}, \epsilon]$ constructs a space $\mathcal{W}$ of the form (3.1) and $T_V \mathbf{w} \in \mathcal{W}$ satisfying

$$\varrho = \| \mathbf{w} - \mathbf{w} \|_{\ell^2(A;V)} \leq \epsilon.$$  

Furthermore, $\dim \mathcal{W}$ is minimal among all spaces of the form (3.1) with $\mathcal{W}_\mu = W^i_\mu$ and satisfying (3.3).

Proof. Equation (3.3) follows from the termination criterion in $\text{Refine}_D$. Convergence is ensured by (3.2) and $W^i_\mu \uparrow V$ for all $\mu$. For the optimality property of the greedy algorithm, we refer to the more general statement [20, Theorem 4.1.5]. □

3.2. Adaptive Galerkin method. Let $\| \cdot \|_A$ denote the energy norm on $\ell^2(A;V)$, i.e., $\| \mathbf{v} \|_A := \sqrt{\langle \mathbf{A} \mathbf{v}, \mathbf{v} \rangle}$. We assume that a routine

$$\text{Galerkin}_{A,f}\{\mathcal{W}, \mathbf{u}_0, \epsilon\} \mapsto [\hat{\mathbf{u}}, \tau]$$

is available which, given a finite-dimensional subspace $\mathcal{W}$ of $L^2_\omega(G;V)$ of the form (1.28), and starting from the initial approximation $\mathbf{u}_0$, iteratively computes $\hat{\mathbf{u}} \in \ell^2(A;V)$ with $T_V \hat{\mathbf{u}} \in \mathcal{W}$ and

$$\| \hat{\mathbf{u}} - \hat{\mathbf{u}} \|_A \leq \tau \leq \epsilon,$$

where $T_V \hat{\mathbf{u}}$ is the Galerkin projection of $\mathbf{u}$ onto $\mathcal{W}$. An example of such a routine, based on a preconditioned conjugate gradient iteration, is given in [22].

We combine the method $\text{Residual}_{A,f}$ for approximating the residual, $\text{Refine}_D$ for refining the Galerkin subspace and $\text{Galerkin}_{A,f}$ for approximating the Galerkin projection, to an adaptive solver $\text{SolveGalerkin}_{A,f}$ similar to [8, 19, 16].

\[
\text{SolveGalerkin}_{A,f}[\epsilon, \gamma, \chi, \vartheta, \omega, \sigma, \alpha, \beta] \mapsto \mathbf{u}_\epsilon
\]

$\mathcal{W}(0) \leftarrow \{0\}$

$\hat{\mathbf{u}}^{(0)} \leftarrow 0$

$\delta_0 \leftarrow \sqrt{(1 - \gamma)^{-1}} \| f \|_{\ell^2(A;V^*)}$

for $k = 0, 1, 2, \ldots$ do

$[\mathbf{w}, \eta, \zeta] \leftarrow \text{Residual}_{A,f}[\epsilon \sqrt{1 - \gamma}, \hat{\mathbf{u}}^{(k)}, \delta_k, \chi, \omega, \alpha, \beta]$

$\delta_k \leftarrow (\eta_k + \zeta_k)/\sqrt{1 - \gamma}$

if $\min(\delta_k, \bar{\delta}_k) \leq \epsilon$ then break

$[\mathcal{W}^{(k+1)}, \bar{\mathbf{u}}^{(k+1)}, \varrho_k] \leftarrow \text{Refine}_D[\mathcal{W}^{(k)}, \mathbf{w}, \sqrt{\eta_k^2 - (\eta_k + \vartheta)/(\eta_k + \zeta_k)^2}]$

$\bar{\varrho}_k \leftarrow (\sqrt{\eta_k^2 - \varrho_k^2 - \zeta_k}/(\eta_k + \zeta_k)$

$[\bar{\mathbf{u}}^{(k+1)}, \tau_k^{(k+1)}] \leftarrow \text{Galerkin}_{A,f}[\mathcal{W}^{(k+1)}, \bar{\mathbf{u}}^{(k)}], \sigma \min(\delta_k, \bar{\delta}_k)]$

$\delta_{k+1} \leftarrow \tau_k^{(k+1)} + \sqrt{1 - \bar{\varrho}_k^2(1 - \gamma)(1 + \gamma)^{-1}} \min(\delta_k, \bar{\delta}_k)$

$\mathbf{u}_\epsilon \leftarrow \bar{\mathbf{u}}^{(k)}$

3.3. Convergence of the adaptive solver. The convergence analysis of the method $\text{SolveGalerkin}_{A,f}$ is based on [8, Lemma 4.1], which generalizes to our vector setting for Galerkin spaces $\mathcal{W}$ of the form (1.28). Let $\Pi_{w}$ denote the orthogonal projection in $\ell^2(A;V)$ onto $T_V^{-1}\mathcal{W}$, and let $\hat{\Pi}_W := \mathbf{D} \Pi_W \mathbf{D}^{-1}$ be the orthogonal projection in $\ell^2(A;V^*)$ onto $\mathbf{D} T_V^{-1}\mathcal{W} = T_V^{-1} \mathcal{W}$. 

\[
\text{SolveGalerkin}_{A,f}[\epsilon, \gamma, \chi, \vartheta, \omega, \sigma, \alpha, \beta] \mapsto \mathbf{u}_\epsilon
\]
Proposition 3.2. Let \( \mathcal{W} \) be as in (3.28), and \( \vartheta \in [0, 1] \). Let \( v \in \mathcal{W} \)

\[
(3.6) \quad \left\| \bar{I}_\mathcal{W}(f - Av) \right\|_{\ell^2(A; V^*)} \geq \vartheta \left\| f - Av \right\|_{\ell^2(A; V^*)}.
\]

Then the Galerkin projection \( \bar{u} \) of \( u \) onto \( \mathcal{W} \) satisfies

\[
(3.7) \quad \| u - \bar{u} \|_A \leq \sqrt{1 - \vartheta^2 \frac{1 - \gamma}{1 + \gamma}} \| u - v \|_A.
\]

Proof. Due to (3.6),

\[
\| u - \bar{u} \|_A \geq \| A \|^{-1/2} \| A(u - v) \|_{\ell^2(A; V^*)} \geq \| A \|^{-1/2} \left\| \bar{I}_\mathcal{W}(f - Av) \right\|_{\ell^2(A; V^*)}
\]

\[
\geq \| A \|^{-1/2} \vartheta \left\| f - Av \right\|_{\ell^2(A; V^*)} \geq \| A \|^{-1/2} \left\| A^{-1} \right\|^{-1/2} \vartheta \| u - v \|_A.
\]

By Galerkin orthogonality,

\[
\| u - \bar{u} \|_A^2 = \| u - v \|_A^2 - \| \bar{u} - v \|_A^2 \leq (1 - \vartheta^2) \| A \|^{-1} \left\| A^{-1} \right\|^{-1} \| u - v \|_A^2.
\]

The assertion follows using the estimates \( \| A \| \leq (1 + \gamma) \) and \( \| A^{-1} \| \leq (1 - \gamma)^{-1} \), which follow from (12.24).

Lemma 3.3. Let \( \epsilon > 0, \chi > 0 \) and \( \alpha, \beta \in (0, 1) \). If \( \vartheta > 0, \omega > 0, \) and \( \omega + \vartheta + \omega \vartheta \leq 1 \), then the space \( \mathcal{W}(k+1) \) in \( \text{SolveGalerkin}_{\mathbf{A}, \mathbf{f}} \) is such that

\[
(3.8) \quad \left\| \bar{I}_{\mathcal{W}(k+1)} r_k \right\|_{\ell^2(A; V^*)} \geq \bar{\vartheta}_k \left\| r_k \right\|_{\ell^2(A; V^*)}
\]

where \( r_k := f - Av^{(k)} \) is the residual at iteration \( k \in \mathbb{N}_0 \), and \( \bar{\vartheta}_k \geq \vartheta \).

Proof. We abbreviate \( z := w_k - u^{(k)} \). Due to \( \zeta_k \leq \omega \eta_k \), the assumption \( \omega + \vartheta + \omega \vartheta \leq 1 \) implies \( \zeta_k + \vartheta (\eta_k + \zeta_k) \leq \eta_k \). Thus the tolerance in \( \text{Refine}_D \) is nonnegative. Since \( u^{(k)} \in \mathcal{W}(k) \subset \mathcal{W}(k+1) \), Proposition 3.1 implies

\[
\vartheta_k = \left\| w_k - w_k \right\|_{\ell^2(A; V)} = \left\| w_k - \Pi_{\mathcal{W}(k+1)} w_k \right\|_{\ell^2(A; V)} = \left\| z - \Pi_{\mathcal{W}(k+1)} z \right\|_{\ell^2(A; V)}.
\]

Consequently,

\[
\left\| \Pi_{\mathcal{W}(k+1)} z \right\|_{\ell^2(A; V)}^2 = \left\| z \right\|_{\ell^2(A; V)}^2 - \left\| z - \Pi_{\mathcal{W}(k+1)} z \right\|_{\ell^2(A; V)}^2 = \eta_k^2 - \vartheta_k^2.
\]

Furthermore, since \( \Pi_{\mathcal{W}(k+1)} \) has norm one, Proposition 2.2 implies

\[
\left\| \Pi_{\mathcal{W}(k+1)} z \right\|_{\ell^2(A; V)} - \left\| \bar{I}_{\mathcal{W}(k+1)} r_k \right\|_{\ell^2(A; V^*)} \leq \left\| \Pi_{\mathcal{W}(k+1)} (z - D^{-1} r_k) \right\|_{\ell^2(A; V)}
\]

\[
\leq \left\| z - D^{-1} r_k \right\|_{\ell^2(A; V)} \leq \zeta_k.
\]

Combining these estimates, we have

\[
\left\| \bar{I}_{\mathcal{W}(k+1)} r_k \right\|_{\ell^2(A; V^*)} \geq \left\| \Pi_{\mathcal{W}(k+1)} z \right\|_{\ell^2(A; V)} - \zeta_k = \sqrt{\eta_k^2 - \vartheta_k^2} - \zeta_k,
\]

and (3.8) follows using \( \| r_k \|_{\ell^2(A; V^*)} \leq \eta_k + \zeta_k \). Finally, \( \eta_k^2 - \vartheta_k^2 \leq (\zeta_k + \vartheta (\eta_k + \zeta_k))^2 \)

implies \( \sqrt{\eta_k^2 - \vartheta_k^2} \geq \zeta_k + \vartheta (\eta_k + \zeta_k) \), and therefore \( \bar{\vartheta}_k = (\sqrt{\eta_k^2 - \vartheta_k^2} - \zeta_k) / (\eta_k + \zeta_k) \geq \vartheta \).

Theorem 3.4. If \( \epsilon > 0, \chi > 0, \vartheta > 0, \omega > 0, \) and \( \omega + \vartheta + \omega \vartheta \leq 1 \), then \( \text{SolveGalerkin}_{\mathbf{A}, \mathbf{f}}(\epsilon, \gamma, \chi, \vartheta, \omega, \sigma, \alpha, \beta) \) constructs a finitely supported \( u_\epsilon \in \ell^2(A; V) \) with

\[
(3.9) \quad \| u - u_\epsilon \|_A \leq \epsilon.
\]
Moreover,

\[ (3.10) \quad \sqrt{1 - \frac{1}{\gamma} - \frac{1}{1 + \gamma}} \delta_k \leq \left\| u - \tilde{u}^{(k)} \right\|_A \leq \min(\delta_k, \tilde{\delta}_k) \]

for all \( k \in \mathbb{N}_0 \) reached by SolveGalerkin\(_A\).f.

**Proof.** Due to the termination criterion of SolveGalerkin\(_A\).f, it suffices to show (3.10). For \( k = 0 \), since \( \|u\|_{L^2(A;V)} \leq \|A^{-1}\|^{1/2} \|u\|_A \),

\[ \left\| u - \tilde{u}^{(0)} \right\|_A^2 = \|u\|_A^2 = \langle f, u \rangle_{L^2(A;V)} \leq \|f\|_{L^2(A;V^*)} \|u\|_{L^2(A;V)} \leq \delta_0 \|u\|_A \]  

Let \( \|u - \tilde{u}^{(k)}\|_A \leq \delta_k \) for some \( k \in \mathbb{N}_0 \). Abbreviating \( r_k := f - A\tilde{u}^{(k)} \), using (1.26) then (2.14), we have

\[ \left\| u - \tilde{u}^{(k)} \right\|_A \leq \frac{1}{\sqrt{1 - \gamma} \|r_k\|_{L^2(A;V^*)}} \leq \frac{\zeta_k + \eta_k}{\sqrt{1 - \gamma}} = \tilde{\delta}_k. \]

If \( \min(\delta_k, \tilde{\delta}_k) \geq \epsilon \), then \( \zeta_k \leq \omega \eta_k \) by Proposition 2.2. Due to Lemma 3.3, Proposition 3.2 implies

\[ \|u - \tilde{u}\|_A \leq \sqrt{1 - \theta^2} \frac{1 - \gamma}{1 + \gamma} \min(\delta_k, \tilde{\delta}_k), \]

where \( \tilde{u} \) is the exact Galerkin projection of \( u \) onto \( \mathcal{W}^{(k+1)} \). By (3.5), \( \tilde{u}^{(k+1)} \) approximates \( \tilde{u} \) up to an error of at most \( \tau_{k+1} \leq \sigma \min(\delta_k, \tilde{\delta}_k) \) in the norm \( \|\cdot\|_A \).

It follows by triangle inequality that

\[ \left\| u - \tilde{u}^{(k+1)} \right\|_A \leq \delta_{k+1}. \]

To show the other inequality in (3.10), we note that for any \( k \in \mathbb{N}_0 \),

\[ \left\| u - \tilde{u}^{(k)} \right\|_A \geq \frac{1}{\sqrt{1 + \gamma}} \|r_k\|_{L^2(A;V^*)} \geq \frac{\eta_k - \zeta_k}{\sqrt{1 + \gamma}} = \sqrt{1 - \frac{\gamma}{\eta_k + \zeta_k}} \tilde{\delta}_k, \]

and \((\eta_k - \zeta_k)(\eta_k + \zeta_k)^{-1} \geq (1 - \omega)(1 + \omega)^{-1} \).

Finally, since

\[ \delta_k \leq \left( \sigma + \sqrt{1 - \theta^2(1 - \gamma)(1 + \gamma)^{-1}} \right)^k \delta_0 \]

and \( \sigma + \sqrt{1 - \theta^2(1 - \gamma)(1 + \gamma)^{-1}} < 1 \) by assumption, the iteration does terminate. \( \square \)

**4. Optimality properties**

**4.1. A semidiscrete algorithm.** The algorithm SolveGalerkin\(_A\).f is derived in Section 3 with arbitrary Galerkin subspaces of the form (1.28). We consider optimality properties of this method in the special case of a single spatial discretization, where a Galerkin subspace \( \mathcal{W} \subset L^2(A;V) \) is fully determined by its set of active indices \( \Xi \subset \Lambda \).

Since the spatial discretization is fixed throughout, only the part of the residual pertaining to the random part of the error needs to be computed to construct refinements. In particular, no adaptive solver is needed to invert \( D \), making this a viable approach if no such solver is available, or whenever only a single spatial discretization is desired. It is not our intent to suggest that such spaces should generally be used in practice. The adaptive method SolveGalerkin\(_A\).f in its full generality has the potential to construct much sparser approximations of \( u \). However, the
heuristic distribution of tolerances in Residual\_A\_f precludes provable optimality statements in this setting; see Remark 2.3.

In this section, we think of the operator \( A \) from (1.1) as being already discretized in space, and \( V \) is, e.g., a finite element space. Thus, abstractly, we consider a semidiscrete version of the algorithm SolveGalerkin\_A\_f.

The Galerkin subspaces \( V^{(k)} \) have the form \( \ell^2(\Xi^{(k)}; V) \) for finite sets \( \Xi^{(k)} \subset \Lambda \). In the subroutine Residual\_A\_f, we assume that Solve\_D inverts \( D \) exactly in \( V \). The parameter \( \alpha \) can thus be set to zero.

In the subsequent refinement step, \( \Xi^{(k)} \) is augmented by sufficiently many elements of \( \text{supp} \ w_k \) to represent \( w_k \) to the desired accuracy. The method Refine\_D reduces to ordering \( \text{supp} \ w_k \) according to \( \| w_k, v \|_V \) and selecting the most important contributions.

In Galerkin\_A\_f, an iterative solver such as a conjugate gradient iteration is used to approximate the Galerkin projection of \( u \) onto \( \ell^2(\Xi^{(k+1)}; V) \). Operations within \( V \) are assumed to be exact.

### 4.2. Optimal choice of subspaces.

For \( v \in \ell^2(A; V) \) and \( N \in \mathbb{N}_0 \), let \( P_N(v) \) be a best \( N \)-term approximation of \( v \), that is, \( P_N(v) \) is an element of \( \ell^2(A; V) \) that minimizes \( \| v - P_N(v) \|_{\ell^2(A; V)} \) over \( P_N \in \ell^2(A; V) \) with \( \# \text{supp} \ v \leq N \). For \( s \in (0, \infty) \), we define

\[
\| v \|_{A^s(A; V)} := \sup_{N \in \mathbb{N}_0} (N + 1)^s \| v - P_N(v) \|_{\ell^2(A; V)}
\]

and

\[
A^s(A; V) := \left\{ v \in \ell^2(A; V) : \| v \|_{A^s(A; V)} < \infty \right\}.
\]

By definition, an optimal approximation in \( \ell^2(A; V) \) of \( v \in A^s(A; V) \) with error tolerance \( \epsilon > 0 \) consists of \( O(\epsilon^{-1/s}) \) nonzero coefficients in \( V \).

For any \( \Xi \subset \Lambda \), let \( P_{\Xi} \) denote the orthogonal projection in \( \ell^2(A; V^*) \) onto \( \ell^2(\Xi; V^*) \). The following statement is adapted from [19, Lemma 2.1] and [16, Lemma 4.1].

**Lemma 4.1.** Let \( \Xi^{(0)} \) be a finite subset of \( \Lambda \) and \( v \in \ell^2(\Xi^{(0)}; V) \). If

\[
0 \leq \gamma < \sqrt{1 - \gamma} \frac{1}{1 + \gamma}
\]

and \( \Xi^{(0)} \subset \Xi^{(1)} \subset \Lambda \) with

\[
\#\Xi^{(1)} \leq \bar{c} \min \left\{ \#\Xi ; \Xi^{(0)} \subset \Xi, \| P_{\Xi}(f - Av) \|_{\ell^2(A; V^*)} \geq \gamma \| f - Av \|_{\ell^2(A; V^*)} \right\}
\]

for a \( \bar{c} \geq 1 \), then

\[
\#(\Xi^{(1)} \setminus \Xi^{(0)}) \leq \bar{c} \min \left\{ \#\hat{\Xi} ; \hat{\Xi} \subset \Lambda, \| u - \hat{u} \|_A \leq \tau \| u - v \|_A \right\}
\]

for \( \tau = \sqrt{1 - \gamma^2(1 + \gamma)(1 - \gamma)^{-1}} \), where \( \hat{u} \) denotes the Galerkin projection of \( u \) onto \( \ell^2(\hat{\Xi}; V) \).

**Proof.** Let \( \hat{\Xi} \) be as in (1.5) and \( \hat{\Xi} := \Xi^{(0)} \cup \hat{\Xi} \). Furthermore, let \( \hat{u} \) and \( \hat{u} \) denote the Galerkin solutions in \( \ell^2(\hat{\Xi}; V) \) and \( \ell^2(\hat{\Xi}; V) \), respectively. Since \( \hat{\Xi} \subset \hat{\Xi} \),
\[ \| \mathbf{u} - \tilde{\mathbf{u}} \|_{\mathcal{A}} \leq \| \mathbf{u} - \tilde{\mathbf{u}} \|_{\mathcal{A}}, \] and by Galerkin orthogonality,
\[ \| \tilde{\mathbf{u}} - \mathbf{v} \|_{\mathcal{A}}^2 = \| \mathbf{u} - \mathbf{v} \|_{\mathcal{A}}^2 - \| \mathbf{u} - \tilde{\mathbf{u}} \|_{\mathcal{A}}^2 \geq (1 - \tau^2) \| \mathbf{u} - \mathbf{v} \|_{\mathcal{A}}^2 = \hat{\theta}^2 \frac{1 + \gamma}{1 - \gamma} \| \mathbf{u} - \mathbf{v} \|_{\mathcal{A}}.
\]

Therefore, using \( \kappa(\mathcal{A}) = \| \mathcal{A} \| \| \mathcal{A}^{-1} \| \leq (1 + \gamma)(1 - \gamma)^{-1} \),
\[ \left\| \Pi_{\Xi}(\mathbf{f} - \mathcal{A}\mathbf{v}) \right\|_{\ell^2(\mathcal{A};\mathcal{V}^\ast)} = \| \mathcal{A}(\tilde{\mathbf{u}} - \mathbf{v}) \|_{\ell^2(\mathcal{A};\mathcal{V}^\ast)} \geq \| \mathcal{A}^{-1} \|^{{-1/2}} \| \tilde{u} - \mathbf{v} \|_{\mathcal{A}} \geq \hat{\theta} \| \mathcal{A} \|^{{1/2}} \| \mathbf{u} - \mathbf{v} \|_{\mathcal{A}} \geq \hat{\theta} \| \mathbf{f} - \mathcal{A}\mathbf{v} \|_{\ell^2(\mathcal{A};\mathcal{V}^\ast)}.
\]

By (4.4), \( \#\Xi^{(1)} \leq \hat{c}\#\Xi \) and, consequently,
\[ \#(\Xi^{(1)} \setminus \Xi^{(0)}) \leq \hat{c}\#(\Xi \setminus \Xi^{(0)}) \leq \hat{c}\#\Xi. \]

We use Lemma 4.1 to show that, under additional assumptions on the parameters, the index sets \( \Xi^{(k)} \) generated by the semidiscrete version of SolveGalerkin\(_{\mathcal{A},f}\) are of optimal size, up to a constant factor.

**Theorem 4.2.** If the conditions of Theorem 3.4 are satisfied,
\[ \hat{\theta} := \frac{\theta(1 + \omega) + 2\omega}{1 - \omega} < \sqrt{\frac{1 - \gamma}{1 + \gamma}}, \]
and \( \mathbf{u} \in \mathcal{A}^s(\mathcal{A};\mathcal{V}) \) for an \( s > 0 \), then for all \( k \in \mathbb{N}_0 \) reached by SolveGalerkin\(_{\mathcal{A},f}\),
\[ \#\Xi^{(k)} \leq 2(\theta/\tau)^{1/s} \left( \frac{(1 + \gamma)(1 + \omega)}{(1 - \gamma)(1 - \omega)} \right)^{1/s} \| \mathbf{u} - \tilde{\mathbf{u}}^{(k)} \|_{\ell^2(\mathcal{A};\mathcal{V})}^{-1/s} \| \mathbf{u} \|_{\mathcal{A}^s(\mathcal{A};\mathcal{V})}^{1/s} \]

with \( \theta = \sigma + \sqrt{1 - \theta^2(1 - \gamma)(1 + \gamma)^{-1}} \) and \( \tau = \sqrt{1 - \theta^2(1 - \gamma)(1 - \gamma)^{-1}} \).

**Proof.** Let \( k \in \mathbb{N}_0, r_k = \mathbf{f} - \mathcal{A}\tilde{\mathbf{u}}^{(k)} \). Also, let \( \varphi = (\varphi_{\nu})_{\nu \in \mathcal{A}}, \varphi_{\nu} := \| w_{k,\nu} - \tilde{u}_{\nu}^{(k)} \|_{\mathcal{V}} \) for the approximation \( w_k - \tilde{u}^{(k)} = (w_{k,\nu} - \tilde{u}_{\nu}^{(k)})_{\nu \in \mathcal{A}} \) of \( D^{-1}r_k \) computed in Residual\(_{\mathcal{A},f}\), and let \( \Delta \subset \text{supp} w_k \) denote the active indices selected by Refine\(_D\).

We note that for \( \alpha := \omega + \vartheta + \omega\vartheta \), we have \( \vartheta = \frac{\alpha - \omega}{1 - \omega} \) and \( \hat{\vartheta} = \frac{\alpha + \omega}{1 + \omega} \). Let \( \Xi^{(k)} \subset \Xi \subset A \) satisfy \( \| \Pi_{\Xi} r_k \|_{\ell^2(\mathcal{A};\mathcal{V}^\ast)} \geq \hat{\vartheta} \| r_k \|_{\ell^2(\mathcal{A};\mathcal{V}^\ast)} \). Also, if \( \tilde{\mathbf{u}}^{(k)} \) is used to refine the discretization, then the tolerance \( \varepsilon \) is not yet reached, and thus \( \| \varphi \|_{\ell^2(\mathcal{A})} - \| r_k \|_{\ell^2(\mathcal{A};\mathcal{V}^\ast)} \leq \omega \| \varphi \|_{\ell^2(\mathcal{A})} \) by Proposition 2.2.

Therefore,
\[ \hat{\vartheta} \| \varphi \|_{\ell^2(\mathcal{A})} \leq \hat{\vartheta} \| r_k \|_{\ell^2(\mathcal{A};\mathcal{V}^\ast)} + \hat{\vartheta} \omega \| \varphi \|_{\ell^2(\mathcal{A})} \leq \| \Pi_{\Xi} r_k \|_{\ell^2(\mathcal{A};\mathcal{V}^\ast)} + \| \Pi_{\Xi} \varphi \|_{\ell^2(\mathcal{A})} \leq \| \Pi_{\Xi} \varphi \|_{\ell^2(\mathcal{A})} \leq \| \Pi_{\Xi} \varphi \|_{\ell^2(\mathcal{A})} \]

and since \( \hat{\vartheta} = (1 + \hat{\vartheta})\omega = \alpha \), it follows that \( \| \Pi_{\Xi} \varphi \|_{\ell^2(\mathcal{A})} \geq \alpha \| \varphi \|_{\ell^2(\mathcal{A})} \). By construction, \( \Delta \) is a set of minimal cardinality with \( \| \Pi_{\Delta} \varphi \|_{\ell^2(\mathcal{A})} \geq \alpha \| \varphi \|_{\ell^2(\mathcal{A})} \) for \( \alpha := \zeta_k \eta_k^{-1} + \vartheta(1 + \zeta_k \eta_k^{-1}) - \alpha \). Consequently, \( \#(\Xi^{(k+1)} \setminus \Xi^{(k)}) \leq \# \Delta \leq \# \Xi \).

Since this holds for any \( \Xi \), using \( \#(\Xi^{(k)}) \leq \Xi \), it follows that
\[ \#\Xi^{(k+1)} \leq 2 \min \left\{ \#\Xi, \Xi \subset \Xi \subset A, \| \Pi_{\Xi} r_k \|_{\ell^2(\mathcal{A};\mathcal{V}^\ast)} \geq \hat{\vartheta} \| r_k \|_{\ell^2(\mathcal{A};\mathcal{V}^\ast)} \right\}.
\]

Lemma 4.1 implies
\[ \#(\Xi^{(k+1)} \setminus \Xi^{(k)}) \leq 2 \min \left\{ \#\Xi, \Xi \subset A, \| \mathbf{u} - \tilde{\mathbf{u}} \|_{\mathcal{A}} \leq \tau \| \mathbf{u} - \tilde{\mathbf{u}}^{(k)} \|_{\mathcal{A}} \right\}.
\]
with \( \tau = \sqrt{1 - \hat{\varphi}^2(1 + \gamma)(1 - \gamma)^{-1}} \), where \( \tilde{u} \) denotes the Galerkin projection of \( u \) onto \( \ell^2(\hat{\Xi}; V) \).

Let \( N \in \mathbb{N}_0 \) be maximal with \( \|u - P_N(u)\|_{\ell^2(A; V)} > \tau(1 + \gamma)^{-1/2} \|u - \hat{u}^{(k)}\|_A \), where \( P_N(u) \) is a best \( N \)-term approximation of \( u \). By (4.1),

\[
N + 1 \leq \|u - P_N(u)\|_{\ell^2(A; V)}^{-1/s} \|u\|_A^{1/s} \leq \tau^{-1/s}(1 + \gamma)^{1/2s} \|u - \hat{u}^{(k)}\|_A^{-1/s} \|u\|_A^{1/s} (A; V).
\]

For \( \Xi_{N+1} := \text{supp} \ P_{N+1}(u) \), by maximality of \( N \),

\[
\|u - \tilde{u}_{N+1}\|_A \leq \|u - P_{N+1}(u)\|_A \leq (1 + \gamma)^{1/2} \|u - P_{N+1}(u)\|_{\ell^2(A; V)} \leq \tau \|u - \hat{u}^{(k)}\|_A
\]

for the Galerkin solution \( \tilde{u}_{N+1} \) in \( \ell^2(\Xi_{N+1}; V) \), and thus

\[
#(\Xi^{(k+1)} \setminus \Xi^{(k)}) \leq 2(N + 1) \leq 2\tau^{-1/s}(1 + \gamma)^{1/2s} \|u - \hat{u}^{(k)}\|_A^{-1/s} \|u\|_A^{1/s} (A; V).
\]

Furthermore, by Theorem 3.4

\[
\|u - \hat{u}^{(k)}\|_A^{-1/s} \leq \left( \sqrt{1 - \gamma \frac{1 - \omega}{1 + \gamma}} \delta_{\bar{j}} \right)^{-1/s}.
\]

We estimate the cardinality of \( \Xi^{(k)} \) by slicing it into increments and applying the above estimates,

\[
#\Xi^{(k)} = \sum_{j=0}^{k-1} #(\Xi^{(j+1)} \setminus \Xi^{(j)}) \leq 2\tau^{-1/s}(1 + \gamma)^{1/2s} \|u\|_A^{1/s} (A; V) \sum_{j=0}^{k-1} \|u - \hat{u}^{(j)}\|_A^{-1/s}
\]

\[
\leq 2 \left( \frac{\tau(1 - \gamma)^{1/2}(1 - \omega)}{(1 + \gamma)(1 + \omega)} \right)^{-1/s} \|u\|_A^{1/s} (A; V) \sum_{j=0}^{k-1} \bar{j}^{-1/s}.
\]

By definition, \( \delta_k \leq \varrho^{k-j} \bar{j} \). Therefore,

\[
\sum_{j=0}^{k-1} \bar{j}^{-1/s} \leq \delta_k^{-1/s} \sum_{j=0}^{k-1} \varrho^{(k-j)/s} = \delta_k^{-1/s} \sum_{i=1}^{k} \varrho^{i/s} = \frac{\varrho^{1/s} \delta_k^{-1/s}}{1 - \varrho^{1/s}}.
\]

The assertion follows using

\[
(1 - \gamma)^{1/2} \left\|u - \hat{u}^{(k)}\right\|_{\ell^2(A; V)} \leq \left\|u - \tilde{u}^{(k)}\right\|_A \leq \delta_k.
\]

**4.3. Complexity estimate.** We first cite an elementary result due to Stechkin connecting the order of summability of a sequence to the convergence of best \( N \)-term approximations in a weaker sequence norm; see e.g. [11, 13]. Note that, although it is formulated only for nonnegative sequences, Lemma 4.3 applies directly to, e.g., Lebesgue–Bochner spaces of Banach space valued sequences by passing to the norms of the elements of such sequences. Also, it applies to sequences with arbitrary countable index sets by choosing a decreasing rearrangement.
Lemma 4.3. Let $0 < p \leq q$ and let $c = (c_n)_{n=1}^{\infty} \in \ell^2$ with $0 \leq c_{n+1} \leq c_n$ for all $n \in \mathbb{N}$. Then
\begin{equation}
\left( \sum_{n=N+1}^{\infty} c_n^q \right)^{1/q} \leq (N+1)^{-r} \|c\|_{\ell^p}, \quad r := \frac{1}{p} - \frac{1}{q} \geq 0
\end{equation}
for all $N \in \mathbb{N}_0$.

Proposition 4.4. Let $s > 0$. If either
\begin{equation}
\|R_m\|_{V \to V^*} \leq s \delta_{R,s}(m+1)^{-s-1} \quad \forall m \in \mathbb{N}
\end{equation}
or
\begin{equation}
\left( \sum_{m=1}^{\infty} \|R_m\|_{V^* \to V}^{s+1} \right)^{1/(s+1)} \leq \delta_{R,s},
\end{equation}
then
\begin{equation}
\|R - R_{[M]}\|_{\ell^2(A;V) \to \ell^2(A;V^*)} \leq \delta_{R,s}(M+1)^{-s} \quad \forall M \in \mathbb{N}_0,
\end{equation}
Proof. By (4.23) and (2.1), using $\|K_m\|_{\ell^2(A) \to \ell^2(A)} \leq 1$,
\begin{equation}
\|R - R_{[M]}\|_{\ell^2(A;V) \to \ell^2(A;V^*)} \leq \sum_{m=M+1}^{\infty} \|R_m\|_{V \to V^*}.
\end{equation}
If (4.9) holds, then (4.11) follows using
\begin{equation}
\sum_{m=M+1}^{\infty} (m+1)^{-s-1} \leq \int_{M+1}^{\infty} t^{-s-1} dt = \frac{1}{s} (M+1)^{-s}.
\end{equation}
If (4.10) is satisfied, then
\begin{equation}
\sum_{m=M+1}^{\infty} \|R_m\|_{V \to V^*} \leq \left( \sum_{m=1}^{\infty} \|R_m\|_{V \to V^*}^{s+1} \right)^{1/(s+1)} (M+1)^{-s}
\end{equation}
by Lemma 4.3.

Remark 4.5. If the assumptions of Proposition 4.4 are satisfied for all $s \in (0,s^*)$, then the operator $R$ is $s^*$-compressible with sparse approximations $R_{[M]}$. In this case, $R$ is a bounded linear map from $\mathcal{A}^s(A;V)$ to $\mathcal{A}^s(A;V^*)$ for all $s \in (0,s^*)$; see [8 Prop. 3.8]. This carries over to the routine $\text{Apply}_R$ in that if $v \in \mathcal{A}^s(A;V)$ and $z$ is the output of $\text{Apply}_R[v,\epsilon]$ for an $\epsilon > 0$, then
\begin{equation}
\# \text{ supp } z \leq \|v\|_{\mathcal{A}^s(A;V)}^{1/s} \epsilon^{-1/s},
\end{equation}
\begin{equation}
\|z\|_{\mathcal{A}^s(A;V^*)} \lesssim \|v\|_{\mathcal{A}^s(A;V)}
\end{equation}
with constants depending only on $s$ and $R$. Moreover, (4.12) is an upper bound for the total number of applications of operators $R_m$ in $\text{Apply}_R[v,\epsilon]$. This follows as in the scalar case (see e.g. [16 Prop. 4.6]), where the additional term $1 + \# \text{ supp } v$ is only due to the approximate sorting of $v$.

We make further assumptions on the routine $\text{RHS}_f$. If $f \in \mathcal{A}^{s}(A;V^*)$ and $\tilde{f}$ is the output of $\text{RHS}_f[\epsilon]$ for an $\epsilon > 0$, then $\tilde{f}$ should satisfy
\begin{equation}
\# \text{ supp } \tilde{f} \lesssim \|f\|_{\mathcal{A}^{s}(A;V^*)}^{1/s} \epsilon^{-1/s}.
\end{equation}
This is clearly satisfied for deterministic \( f \), and is achieved for the right-hand sides of the form \( R \mathbf{w} \) for a finitely supported \( \mathbf{w} \), stemming for example from inhomogeneous essential boundary conditions, by using \( \text{Apply}_R \) to approximate this product. Note that if \( \mathbf{u} \in A^s(\mathbf{A};V) \) and \( R \) is \( s^* \)-compressible with \( s < s^* \), then also \( A \) is \( s^* \)-compressible, and therefore \( \| \mathbf{f} \|_{A^s(\mathbf{A};V^*)} \lesssim \| \mathbf{u} \|_{A^s(\mathbf{A};V)} \).

**Lemma 4.6.** Under the conditions of Theorem 4.2, \( \| \mathbf{u}^{(k)} \|_{A^s(\mathbf{A};V)} \leq C \| \mathbf{u} \|_{A^s(\mathbf{A};V)} \quad \forall k \in \mathbb{N}_0 \), with
\[
C = 1 + \frac{2^{1+s} \varrho(1+\gamma)(1+\omega)}{\tau(1-\varrho^{1/s})^s(1-\gamma)(1-\omega)},
\]
for \( \varrho = \sigma + \sqrt{1-\varrho^2(1-\gamma)(1+\gamma)^{-1}} \) and \( \tau = \sqrt{1-\varrho^2(1+\gamma)(1-\gamma)^{-1}} \).

**Proof.** Let \( k \in \mathbb{N}_0 \). For any \( N \geq \# \Xi^{(k)} \), \( \| \tilde{\mathbf{u}}^{(k)} - P_N(\tilde{\mathbf{u}}^{(k)}) \|_{\ell^2(\mathbf{A};V)} = 0 \). For \( N \leq \# \Xi^{(k)} - 1 \),
\[
\begin{align*}
\| \tilde{\mathbf{u}}^{(k)} - P_N(\tilde{\mathbf{u}}^{(k)}) \|_{\ell^2(\mathbf{A};V)} &\leq \| \tilde{\mathbf{u}}^{(k)} - \Pi_{\Xi_u} \tilde{\mathbf{u}}^{(k)} \|_{\ell^2(\mathbf{A};V)} \\
&\leq \| \mathbf{u} - \Pi_{\Xi_u} \mathbf{u} \|_{\ell^2(\mathbf{A};V)} + 2 \| \mathbf{u} - \tilde{\mathbf{u}}^{(k)} \|_{\ell^2(\mathbf{A};V)},
\end{align*}
\]
where \( \Xi_u := \text{supp} P_N(\mathbf{u}) \), such that \( \Pi_{\Xi_u} \mathbf{u} = P_N(\mathbf{u}) \) and
\[
\| \mathbf{u} - \Pi_{\Xi_u} \mathbf{u} \|_{\ell^2(\mathbf{A};V)} \leq (N+1)^{-s} \| \mathbf{u} \|_{A^s(\mathbf{A};V)} .
\]
Furthermore, Theorem 4.2 implies
\[
\begin{align*}
\| \mathbf{u} - \tilde{\mathbf{u}}^{(k)} \|_{\ell^2(\mathbf{A};V)} &\leq \frac{2^s \varrho(1+\gamma)(1+\omega)}{\tau(1-\varrho^{1/s})^s(1-\gamma)(1-\omega)} (\# \Xi^{(k)})^{-s} \| \mathbf{u} \|_{A^s(\mathbf{A};V)} ,
\end{align*}
\]
and \((N+1)^{-s} \leq (\# \Xi^{(k)})^s \) by the definition of \( N \). Consequently,
\[
\| \tilde{\mathbf{u}}^{(k)} \|_{A^s(\mathbf{A};V)} = \sup_{N \in \mathbb{N}_0} (N+1)^{-s} \| \tilde{\mathbf{u}}^{(k)} - P_N(\tilde{\mathbf{u}}^{(k)}) \|_{\ell^2(\mathbf{A};V)} \leq C \| \mathbf{u} \|_{A^s(\mathbf{A};V)}
\]
with \( C \) from (4.16). \( \square \)

**Theorem 4.7.** Let the conditions of Theorem 4.2 be satisfied. If (4.14) and the assumptions of Proposition 4.4 hold for all \( s \in (0,s^*) \), then for any \( \epsilon > 0 \) and any \( s \in (0,s^*) \), the total number of applications of \( D \), \( A^s_{\nu} \), and \( D^{-1} \) in SolveGalerkin_{A,f}[\epsilon,\gamma,\chi,\vartheta,\omega,\sigma,0,\beta] \) is bounded by \( \| \mathbf{u} \|_{A^s(\mathbf{A};V)}^{1/s} \) up to a constant factor depending only on the input arguments other than \( \epsilon \). The same bound holds for the total number of applications of \( R_m, m \in \mathbb{N} \), up to an additional factor of \( \max_{\nu} \mu \in \text{supp} \mu \), \# \text{supp} \mu.

**Proof.** Let \( k \in \mathbb{N}_0 \); we consider the \( k \)-th iteration of the loop in SolveGalerkin_{A,f}. The routine Residual_{A,f}[\epsilon \sqrt{1-\gamma},\tilde{\mathbf{u}}^{(\Xi^{(k)})},\delta_k,\chi,\omega,\beta] \) begins with \# \Xi^{(k)} applications of \( D \). Due to the geometric decrease in tolerances, the complexity of the loop in Residual_{A,f} is dominated by that of its last iteration. By Remark 4.5 and Lemma 4.6, up to a constant factor, the number of applications of \( D^{-1} \) and \( R_m \) is bounded by \( \| \mathbf{u} \|_{A^s(\mathbf{A};V)}^{1/s} \zeta_k^{-1/s} \), and \( \zeta_k \geq \delta_k \).
Next, assuming the termination criterion of \texttt{SolveGalerkin}_{\mathbf{A},f} is not satisfied, the routine \texttt{Galerkin}_{\mathbf{A},f}[\Xi^{(k+1)}, \mathbf{w}, \sigma \min(\delta_k, \tilde{\delta}_k)] is called to iteratively approximate the Galerkin projection onto \(\ell^2(\Xi^{(k+1)}; V)\). Since only a fixed relative error reduction is required, the number of iterations remains bounded. Therefore, the number of applications of \(D^{-1}\) and \(A_{\nu \nu}\) is bounded by \(#\Xi^{(k+1)}\) and the total number of applications of \(R_m\), \(m \in \mathbb{N}\), is bounded by \(2\lambda(\Xi^{(k+1)})#\Xi^{(k+1)}\), where \(\lambda(\Xi^{(k+1)})\) denote the average length of indices in \(\Xi^{(k+1)}\); see [22, Proposition 3.5]. Since the sets \(\Xi^{(k)}\) are nested, \(\lambda(\Xi^{(k+1)}) \leq \max_{m \in \supp \mathbf{u}} \mu, \sup \mu\). Furthermore, by Theorems 3.4 and 4.2, \(\sum_{i=0}^{k-1} \tilde{\delta}_j^{1/s} \leq \| \mathbf{u} \|_{A^s(A; V)}^{1/s} \delta^{1/s}_{k-1} \).

Let \(k\) be such that \(\mathbf{u}_* = \tilde{\mathbf{u}}^{(k)}\). Due to the different termination criterion, the complexity of the last call of \texttt{Residual}_{\mathbf{A},f} can be estimated by \(\| \mathbf{u} \|_{A^s(A; V)}^{1/s} \zeta^{1/s}_{k-1}\) with \(\zeta_k \geq \epsilon\). This bound obviously also holds for \#\Xi^{(k)}\), and thus for the complexity of the final call of \texttt{Galerkin}_{\mathbf{A},f}.

Combining all of the above estimates, the number of applications of \(D^{-1}, D, A_{\nu \nu}\) and \(R_m\), \(m \in \mathbb{N}\), in \texttt{SolveGalerkin}_{\mathbf{A},f} is bounded by

\[
\| \mathbf{u} \|_{A^s(A; V)}^{1/s} \left( \epsilon^{-1/s} + \sum_{j=0}^{k-1} \tilde{\delta}_j^{-1/s} \right) .
\]

Furthermore, \(\tilde{\delta}_{k-1} \geq \epsilon\), and using \(\delta_{k-1} \leq \tilde{\delta}_{k-1}^{-1/\gamma} \tilde{\gamma}_j\),

\[
\sum_{j=0}^{k-2} \tilde{\delta}_j^{-1/s} \leq \tilde{\delta}_{k-1}^{-1/s} \sum_{j=0}^{k-2} \tilde{\delta}_{k-1}^{-1/(1-\gamma)} = \tilde{\delta}_{k-1}^{-1/s} \sum_{i=1}^{k-1} \tilde{\delta}_i^{1/s} \leq \tilde{\delta}_{k-1}^{-1/s} \frac{\tilde{\delta}_i^{1/s}}{1 - \tilde{\delta}_i^{1/s}},
\]

where \(\tilde{\gamma} = \sigma + \sqrt{1 - \tilde{\gamma}^2(1 - \gamma)(1 + \gamma)^{-1}} < 1\). The assertion follows since \(\tilde{\delta}_{k-1} \geq \epsilon\).

## 5. Computational examples

### 5.1. Application to isotropic diffusion

We consider the isotropic diffusion equation on a bounded Lipschitz domain \(G \subset \mathbb{R}^d\) with homogeneous Dirichlet boundary conditions. For any uniformly positive \(a \in L^\infty(G)\) and any \(f \in L^2(G)\), we have

\[
-\nabla \cdot (a(x)\nabla u(x)) = f(x), \quad x \in G,
\]

\[
u(x) = 0, \quad x \in \partial G.
\]

We view \(f\) as fixed, but allow \(a\) to vary, giving rise to a parametric operator

\[
A_0(a) : H^1_0(G) \to H^{-1}(G), \quad v \mapsto -\nabla \cdot (a \nabla v),
\]

which depends continuously on \(a \in L^\infty(G)\).

We model the coefficient \(a\) as a bounded random field, which we expand as a series

\[
a(y, x) := \bar{a}(x) + \sum_{m=1}^{\infty} y_m a_m(x).
\]

Since \(a\) is bounded, \(a_m\) can be scaled such that \(y_m \in [-1, 1]\) for all \(m \in \mathbb{N}\). Therefore, \(a\) depends on a parameter \(y = (y_m)_{m=1}^{\infty} \in \Gamma = [-1, 1]^{\infty}\).
We define the parametric operator \( A(y) := A_0(a(y)) \) for \( y \in \Gamma \). Due to the linearity of \( A_0 \),

\[
A(y) = D + R(y), \quad R(y) := \sum_{m=1}^{\infty} y_m R_m \quad \forall y \in \Gamma
\]

with convergence in \( \mathcal{L}(H^1_0(G), H^{-1}(G)) \), for

\[
D := A_0(\bar{a}): H^1_0(G) \to H^{-1}(G), \quad v \mapsto -\nabla \cdot (\bar{a} \nabla v),
\]

\[
R_m := A_0(a_m): H^1_0(G) \to H^{-1}(G), \quad v \mapsto -\nabla \cdot (a_m \nabla v), \quad m \in \mathbb{N}.
\]

To ensure bounded invertibility of \( D \), we assume there is a constant \( \delta > 0 \) such that

\[
\text{ess inf}_{x \in G} \bar{a}(x) \geq \delta^{-1}.
\]

We refer, e.g., to \([22, 20, 26]\) for further details.

5.2. A posteriori error estimation. Let the spaces \( W_{\nu} \) from Section 1.5 be finite element spaces of continuous, piecewise smooth functions on meshes \( \mathcal{T}_\nu \) which contain at least the piecewise linear functions on \( \mathcal{T}_\nu \). We assume that these meshes are compatible in the sense that for any \( T_\mu \in \mathcal{T}_\mu \) and \( T_\nu \in \mathcal{T}_\nu \), the intersection \( T_\mu \cap T_\nu \) is either empty, equal to \( T_\mu \), or equal to \( T_\nu \). We denote the set of faces of \( \mathcal{T}_\nu \) by \( \mathcal{F}_\nu \) and define \( h_T \) and \( h_F \) as the diameters of \( T \in \mathcal{T}_\nu \) and \( F \in \mathcal{F}_\nu \), respectively.

In Residual, a generic solver Solve\(_D\) is used to approximate \( D^{-1}g_\nu \) up to a prescribed tolerance. In the present finite element setting, this requires a reliable a posteriori error estimator to verify that the desired accuracy is attained.

The vector \( g = (g_\nu)_{\nu \in \Lambda} \) is the approximation of \( f - Rv \) computed with RHS\(_f\) and Apply\(_R\). For the call of Residual, inside SolveGalerkin, \( v \) is the approximate solution \( \tilde{u}^{(k)} \). Thus \( g_\nu \) has the form

\[
g_\nu = \tilde{f}_\nu - \sum_{i=1}^{k} \kappa_i R_m i, v_i,
\]

where \( \tilde{f}_\nu \) is the approximation of \( f_\nu \) generated by RHS\(_f\), \( v_i = v_{\mu_i} \) for some \( \mu_i = \nu \pm \epsilon_{m_i} \) selected by Apply\(_R\), and \( \kappa_i \) refer to the constants \( \alpha_{m_i}^{m_i} \) and \( \beta_{m_i}^{m_i} \) from (1.22). We abbreviate \( \mathcal{T}_\nu := \mathcal{T}_\mu \).

Standard error estimators have difficulties on faces of \( \mathcal{T}_\nu \) that are not in the skeleton of \( \mathcal{T}_\nu \), since \( g_\nu \) is singular on these faces. For all \( i \), let \( \tilde{v}_i \) be an approximation of \( v_i \) that is piecewise smooth on \( \mathcal{T}_\nu \). Replacing \( g_\nu \) by

\[
\tilde{g}_\nu := \tilde{f}_\nu - \sum_{i=1}^{k} \kappa_i R_m i, \tilde{v}_i
\]

induces an error

\[
\|D^{-1}g_\nu - D^{-1}\tilde{g}_\nu\|_V \leq \sum_{i=1}^{k} |\kappa_i| \left\| \frac{a_m}{\bar{a}} \right\|_{L^\infty(G)} \|v_i - \tilde{v}_i\|_V =: \text{EST}_\nu^D
\]

since

\[
\sup_{\|z\|_V = 1} \left| \int_G a_m \nabla v \cdot \nabla z \, dx \right| \leq \left\| \frac{a_m}{\bar{a}} \right\|_{L^\infty(G)} \sup_{\|z\|_V = 1} \int_G |\bar{a} \nabla v \cdot \nabla z| \, dx = \left\| \frac{a_m}{\bar{a}} \right\|_{L^\infty(G)} \|v\|_V
\]
for all \( m \in \mathbb{N} \) and all \( v \in H_0^1(G) \).

Let \( \bar{w}_\nu \in W_\nu \) be the Galerkin projection of \( D^{-1}\bar{g}_\nu \), i.e.,

\[
(5.9) \quad \int_G \bar{a} \nabla \bar{w}_\nu \cdot \nabla z \, dx = \int_G \tilde{f}_\nu z \, dx - \sum_{i=1}^{k} \kappa_i \int_G a_{m_i} \nabla \bar{v}_i \cdot \nabla z \, dx \quad \forall z \in W_\nu .
\]

Abbreviating

\[
\sigma_\nu := \bar{a} \nabla \bar{w}_\mu + \sum_{i=1}^{k} \kappa_i a_{m_i} \nabla \bar{v}_i ,
\]

the residual of \( \bar{w}_\nu \) is the functional

\[
(5.11) \quad r_\nu(\bar{w}_\nu; z) = \int_G \bar{g}_\nu z - \bar{a} \nabla \bar{w}_\nu \cdot \nabla z \, dx = \int_G \tilde{f}_\nu z - \sigma_\nu \cdot \nabla z \, dx , \quad z \in H_0^1(G) .
\]

Due to the Riesz isomorphism,

\[
(5.12) \quad \| D^{-1}\bar{g}_\nu - \bar{w}_\nu \|_V = \sup_{z \in H_0^1(G) \setminus \{0\}} \frac{|r_\nu(\bar{w}_\nu; z)|}{\| z \|_V} \leq \sqrt{\delta} \sup_{z \in H_0^1(G) \setminus \{0\}} \frac{|r_\nu(\bar{w}_\nu; z)|}{\| z \|_{H^1(G)}},
\]

with \( \delta \) from (5.5).

For all \( T \in \mathfrak{T}_\nu \), let

\[
(5.13) \quad R_{\nu,T}(\bar{w}_\nu) := h_T \left\| \tilde{f}_\nu + \nabla \cdot \sigma_\nu \right\|_{L^2(T)} ,
\]

where the dependence on \( \bar{w}_\nu \) is implicit in \( \sigma_\nu \). Also, let

\[
(5.14) \quad R_{\nu,F}(\bar{a}_\nu) := h_F^{1/2} \left\| [\sigma_\nu] \right\|_{L^2(F)} ,
\]

where \([·]\) is the normal jump over the face \( F \in \mathfrak{F}_\nu \). These terms combine to

\[
(5.15) \quad \text{EST}^R_\nu(\bar{w}_\nu) := \left( \sum_{T \in \mathfrak{T}_\nu} R_{\nu,T}(\bar{w}_\nu)^2 + \sum_{F \in \mathfrak{F}_\nu} R_{\nu,F}(\bar{a}_\nu)^2 \right)^{1/2}.
\]

The following statement is a straightforward adaptation of the standard result from, e.g., [29, 25, 1] on reliability of residual error estimators.

**Theorem 5.1.** For all \( z \in H_0^1(G) \),

\[
(5.16) \quad |r_\nu(\bar{w}_\nu; z)| \leq C \, \text{EST}^R_\nu(\bar{w}_\nu) \| z \|_{H^1(G)}
\]

with a constant \( C \) depending only on the shape regularity of \( \mathfrak{T}_\nu \).

**Corollary 5.2.** The Galerkin projection \( \bar{w}_\nu \) from (5.9) satisfies

\[
(5.17) \quad \| D^{-1}\bar{g}_\nu - \bar{w}_\nu \|_V \leq \text{EST}^P_\nu + \sqrt{\delta} C \, \text{EST}^R_\nu(\bar{w}_\nu)
\]

for \( \delta \) from (5.5) and \( C \) from Theorem 5.1.

**Proof.** The assertion follows by triangle inequality using (5.8), (5.12) and (5.16). \( \square \)
5.3. **Numerical computations.** We consider as a model problem the diffusion equation (5.1) on the one-dimensional domain $G = (0, 1)$. For two parameters $k$ and $\gamma$, the diffusion coefficient has the form

\[ a(y, x) = 1 + \frac{1}{c} \sum_{m=1}^{\infty} y_m \frac{1}{m^k} \sin(m\pi x), \quad x \in (0, 1), \quad y \in \Gamma = [-1, 1] \]

where $c$ is chosen as

\[ c = \gamma \sum_{m=1}^{\infty} \frac{1}{m^k}, \]

such that $|a(y, x) - 1|$ is always less than $\gamma$. For the distribution of $y \in \Gamma$, we consider the countable product of uniform distributions on $[-1, 1]$; the corresponding family of orthonormal polynomials is the Legendre polynomial basis.

In all of the following computations, the parameters are $k = 2$ and $\gamma = 1/2$. A few realizations of $a(y)$ and the resulting solutions $u(y)$ of (5.1) are plotted in Figure 1.

![Figure 1. Realizations of $a(y, x)$ (left) and $u(y, x)$ (right).](image)

The parameters of SolveGalerkinAf are set to $\chi = 1/8$, $\vartheta = 0.57$, $\omega = 1/4$, $\sigma = 0.01114$, $\alpha = 1/20$ and $\beta = 0$. These values do not satisfy the assumptions of Theorem 4.2; however, the method executes substantially faster than with parameters for which the theorem applies. All computations were performed in Matlab on a workstation with an AMD Athlon™ 64 X2 5200+ processor and 4GB of memory.

We consider a multilevel discretization in which the a posteriori error estimator from Section 5.2 is used to determine an appropriate discretization level independently for each coefficient. A discretization level $j_\mu$, which represents linear finite elements on a uniform mesh with $2^{j_\mu}$ cells, is assigned to each index $\mu$ with the goal of equidistributing the estimated error among all coefficients. In particular, different refinement levels are used to approximate different coefficients $u_\mu$.

In Figure 2 on the left, the errors are plotted against the number of degrees of freedom, which refers to the total number of basis functions used in the discretization, i.e., the sum of $2^{j_\mu} - 1$ over all $\mu$. On the right, we plot the errors against an estimate of the computational cost. This estimate takes scalar products, matrix-vector multiplications and linear solves into account. The total number of each of these operations on each discretization level is tabulated during the computation, weighted by the number of degrees of freedom on the discretization level,
We compare the discretizations generated adaptively by SolveGalerkin\(_{AF}\) with the heuristic a priori adapted sparse tensor product construction from [5]. Using the notation of [26, Section 4], we set \(\gamma = 2\) and \(\eta_m = 1/(r_m + \sqrt{1 + r_m^2})\) for \(r_m = cm^2/2\) and \(c\) from [5, 19]. These values are similar to those used in the computational examples of [5]. The coarsest spatial discretization used in the sparse tensor product contains 16 elements.

In order to isolate the stochastic discretization, we also consider a fixed spatial discretization, using linear finite elements on a uniform mesh of \((0, 1)\) with 1024 elements to approximate all coefficients. This mesh is sufficiently fine such that the finite element error is negligible compared to the total error. We refer to these simpler versions of the numerical methods as single level discretizations.
Figure 4. Slices of index sets generated by \texttt{SolveGalerkin}_A,f (left) and [5] (right) with single level discretization (top) and multilevel discretization (bottom). All sets correspond to the rightmost points in Figure 3. Active indices with support in \{1, 2\} are plotted; the level of the finite element discretization is proportional to the radius of the circle.

The single level versions of \texttt{SolveGalerkin}_A,f and the sparse tensor method construct discretizations of equal quality, with only a slight advantage for the adaptive algorithm. However, with a multilevel discretization, \texttt{SolveGalerkin}_A,f converges faster than the sparse tensor method, with respect to the number of degrees of freedom. At least in this example, the adaptively constructed discretizations are more efficient than sparse tensor products.

As index sets \(\Xi \subset \Lambda\) are infinite dimensional in the sense that they can contain indices of arbitrary length, they are difficult to visualize in only two dimensions. In Figure 4, we plot two-dimensional slices of sets generated by \texttt{SolveGalerkin}_A,f and the sparse tensor construction from [5]. We consider only those indices which are zero in all dimensions after the second, and plot their values in the first two dimensions. The upper plots depict index sets generated using single level discretizations; dots refer to active indices. The lower plots illustrate the discretizations generated with multilevel finite element discretizations. The radii of the circles are proportional to the discretization level.

The bottom two plots in Figure 4 illustrate differences between the discretizations generated by \texttt{SolveGalerkin}_A,f and the sparse tensor construction. The former has many fewer active indices, but higher discretization levels for some of these. For example, the coefficient of the constant polynomial is approximated on meshes with 4096 and 256 elements, respectively. Also, while the sets constructed by sparse tensorization appear triangular in this figure, the adaptively generated index sets are somewhat more convex. All of the sets are anisotropic in the sense that the first dimension is discretized more finely than the second.

We use the convergence curves in Figures 2 and 3 to empirically determine convergence rates of \texttt{SolveGalerkin}_A,f. The convergence rate with respect to the
total number of degrees of freedom is $2/3$, which is faster than the approximation rate of $1/2$ shown in [11, 10]. It also compares favorably to the sparse tensor construction, which converges with rate $1/2$. However, when considering convergence with respect to the computational cost, the rate of SolveGalerkin reduces to $1/2$ also. We suspect that this is due to the approximation of the residual, which is performed on a larger set of active indices than the subsequent approximation of the Galerkin projection.

For the case of a single finite element mesh, [11, 10] show an approximation rate of $3/2$, whereas we observe a rate of $1$ for both SolveGalerkin and sparse tensorization. In principle, it is possible that SolveGalerkin does not converge with the optimal rate in this example, since the parameters used in the computations do not satisfy the assumptions of Theorem 4.2. Alternatively, due to large constants in the approximation estimates, the asymptotic rate may not be perceivable for computationally accessible tolerances.

**Conclusion**

The adaptive method SolveGalerkin efficiently constructs Galerkin spaces and approximations of the corresponding Galerkin projections for elliptic boundary value problems with random coefficients. It is proven to converge, and provides a reliable and efficient bound for the mean square error. In the case of a fixed spatial discretization, the Galerkin subspaces are shown to be optimal, and the algorithm has linear complexity with respect to the number of active polynomial modes, up to a logarithmic term in the computation of the Galerkin projection.

This solver has a modular structure, which allows any discretization of the spatial domain. For a model problem, we consider finite elements with a residual-based a posteriori error estimator. A minor modification of standard estimators is needed to account for finite element functions in the source term.

Numerical computations show that adaptively computed approximate solutions can be sparser than a sparse tensor product construction. Convergence with respect to the total number of degrees of freedom or the total computational cost agrees with or surpasses approximation estimates shown by nonconstructive means in the case of a multilevel spatial discretization.

**References**


Seminar for Applied Mathematics, ETH Zurich, Rämistrasse 101, CH-8092 Zurich, Switzerland

Current address: Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, Indiana 47907

E-mail address: cgittels@purdue.edu