Numer. Math. (2014) 126:471–513
DOI 10.1007/s00211-013-0572-2
Numerische Mathematik

Adaptive wavelet methods for elliptic partial differential equations with random operators

Claude Jeffrey Gittelson

Received: 27 May 2011 / Revised: 20 March 2013 / Published online: 17 July 2013

Abstract We apply adaptive wavelet methods to boundary value problems with random coefficients, discretized by wavelets in the spatial domain and tensorized polynomials in the parameter domain. Greedy algorithms control the approximate application of the fully discretized random operator, and the construction of sparse approximations to this operator. We suggest a power iteration for estimating errors induced by sparse approximations of linear operators.

Mathematics Subject Classification (2000) 35R60 · 47B80 · 60H35 · 65C20 · 65N12 · 65N22 · 65J10 · 65Y20

1 Introduction

Uncertain coefficients in boundary value problems can be modeled as random variables or random fields. Stochastic Galerkin methods approximate the solution of the resulting random partial differential equation 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, such as a polynomial chaos basis, see [1,18,21,28,37–40].

C. J. Gittelson (✉)
Seminar for Applied Mathematics, ETH Zurich, Rämistrasse 101, 8092 Zurich, Switzerland
e-mail: claude.gittelson@sam.math.ethz.ch

Present address:
C. J. Gittelson
Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA

The primary obstacle in applying these methods is the construction of suitable spaces in which to compute an approximate solution. Sparse tensor product constructions have been shown to be highly effective in [5,6,31,36]. Given sufficient prior knowledge on the regularity of the solution, these methods can be tuned to achieve nearly optimal complexity.

An adaptive approach, requiring less prior information, has been studied in [24, 26,27]; see also e.g. [12] for complementary regularity results, and [7] for a similar approach for stochastic loading instead of a random operator. These methods use techniques from the adaptive wavelet algorithms [9,10,22] to select active polynomial chaos modes. Each of these is a deterministic function, and is approximated e.g. by adaptive finite elements.

Although these methods perform well in a model problem, the suggested equidistribution of error tolerances among all active polynomial chaos modes is only a heuristic. The theoretical analysis of these methods currently does not guarantee optimal convergence with respect to the full stochastic and spatial discretization.

In the present work, we apply adaptive wavelet methods simultaneously to the stochastic and spatial bases, omitting the intermediate semidiscrete approximation stage. This takes full advantage of the adaptivity in these methods, and in particular their celebrated optimality properties apply to the fully discretized stochastic equation.

We provide an overview of adaptive wavelet methods for general bi-infinite discrete positive symmetric linear systems in Sect. 2, including convergence analysis and optimality properties. The particular algorithm we present is based on [9,20,22], where refinements are based on approximations of the residual. We suggest a new updating procedure for the tolerance in the computation of the residual that ensures a geometric decrease in the tolerance while simultaneously preventing this tolerance from becoming unnecessarily small.

Section 3 discusses a greedy algorithm for a class of optimization problems that appear in certain subroutines of our adaptive algorithm. In Sect. 4, this greedy algorithm is used within a generic adaptive application routine for s^*-compressible linear operators. Apart from the introduction of the greedy method, this routine and its analysis in Sect. 5 are based primarily on [20]. The concepts of s^*-compressibility and s^*-computability are reviewed in Sect. 4.1.

This efficient approximate application hinges on a sequence of sparse approximations to the discrete operator, and uses estimates of their respective errors. Although convergence rates for such approximations have been shown e.g. in [34], explicit error bounds do not seem to be available. In Sect. 6, we consider a power method for approximating these errors in the operator norm. We provide an analysis of an idealized method, and suggest a practical variant using some ideas from adaptive wavelet methods. This is different from [17] and references therein, where the smallest eigenvalue of e.g. a discretized differential operator is computed by an inverse iteration, in that we do not assume a discrete spectrum, and thus do not approximate an eigenvector, and in that we compute the maximum of the spectrum rather than the minimum.

Random operator equations and their discretization by tensorized polynomials on the parameter domain and a Riesz basis of the underlying Hilbert space are presented in

Sect. 7. Although our discussion is limited to positive symmetric systems for simplicity, all statements extend to nonsymmetric linear systems, and the adaptive algorithm applies to these by passing to the normal equations as in [10]. Similarly, the Riesz basis could be replaced by frames of the domain and codomain of the operator, and complex Hilbert spaces pose no additional difficulties.

In Sect. 8, we construct a sequence of sparse approximations of the discrete random operator. This again makes use of a greedy algorithm. Section 8 discusses the abstract properties of s^*-compressibility and s^*-computability for this operator, which are used in the analysis of the adaptive application routine.

Finally, in Sect. 10, we present a brief example to illustrate our results. We compare the expected s^*-compressibility to approximation rates from [12]. The smaller of these determines the efficiency of adaptive wavelet methods applied to random boundary value problems.

Throughout the paper, $\mathbb{N}_0$ denotes the set of natural numbers including zero and $\mathbb{N} := \mathbb{N}_0 \backslash \{0\}$. Furthermore, the notation $x \lesssim y$ is an abbreviation for $x \leq Cy$ with a generic constant C; $\mathscr{L}(X, Y)$ denotes the space of bounded linear from X to Y, endowed with the operator norm $\|\cdot\|_{X\to Y}$, and we use the abbreviation $\mathscr{L}(X) := \mathscr{L}(X, X)$.

2 Adaptive wavelet methods

2.1 An adaptive Galerkin solver

We consider a bounded linear operator $\mathbf{A} \in \mathscr{L}(\ell^2)$, which we interpret also as a bi-infinite matrix. For simplicity, we consider the index sets in the domain and codomain to be $\mathbb{N}$, although we will later tacitly substitute other countable sets.

We assume that $\mathbf{A}$ is positive symmetric and boundedly invertible, and consider the equation

$$\mathbf{Au} = \mathbf{f} \tag{2.1}$$

for a $\mathbf{f} \in \ell^2$. Let $\|\cdot\|_{\mathbf{A}}$ denote the norm on ℓ^2 induced by $\mathbf{A}$, which we will refer to as the energy norm.

We briefly discuss a variant of the adaptive solver from [9,20,22] for (2.1). This method selects a nested sequence of finite sections of the infinite linear system, and solves these to appropriate tolerances. In each step, an approximation of the residual is computed in order to estimate the error and, if necessary, enlarge the set of active indices. For extensions of this method and alternative approaches, we refer to [9,10,14–16,32,33] and the survey [35].

We assume that the action of $\mathbf{A}$ can be approximated by a routine

$$\mathtt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon] \mapsto \mathbf{z}\ , \quad \|\mathbf{Av} - \mathbf{z}\|_{\ell^2} \leq \varepsilon\ , \tag{2.2}$$

for finitely supported vectors $\mathbf{v}$. Similarly, we require a routine

$$\texttt{RHS}_{\mathbf{f}}[\varepsilon] \mapsto \mathbf{g}\,, \quad \|\mathbf{f}-\mathbf{g}\|_{\ell^2} \leq \varepsilon\,, \tag{2.3}$$

to approximate the right hand side $\mathbf{f}$ of (2.1) to an arbitrary precision ε. These building blocks are combined in $\texttt{Residual}_{\mathbf{A},\mathbf{f}}$ to compute the residual up to an arbitrary relative error.

$\texttt{Residual}_{\mathbf{A},\mathbf{f}}[\varepsilon, \mathbf{v}, \eta_0, \chi, \omega, \beta] \mapsto [\mathbf{r}, \eta, \zeta]$

$\zeta \longleftarrow \chi\eta_0$
repeat
  $\mathbf{r} \longleftarrow \texttt{RHS}_{\mathbf{f}}[\beta\zeta] - \texttt{Apply}_{\mathbf{A}}[\mathbf{v}, (1-\beta)\zeta]$
  $\eta \longleftarrow \|\mathbf{r}\|_{\ell^2}$
  if $\zeta \leq \omega\eta$ *or* $\eta + \zeta \leq \varepsilon$ **then break**
  $\zeta \longleftarrow \omega\frac{1-\omega}{1+\omega}(\eta+\zeta)$

Remark 2.1 The loop in $\texttt{Residual}_{\mathbf{A},\mathbf{f}}$ terminates either if the residual is guaranteed to be smaller than ε, or if the tolerance ζ in the computation of the residual is less than a constant fraction ω of the approximate residual. If neither criterion is met, since $\zeta > \omega\eta$, the updated tolerance satisfies

$$\omega(\eta-\zeta) < \omega\frac{1-\omega}{1+\omega}(\eta+\zeta) < (1-\omega)\zeta. \tag{2.4}$$

This ensures a geometric decrease of ζ while also preventing ζ from becoming unnecessarily small. Indeed, since $\eta+\zeta$ and $\eta-\zeta$ are upper and lower bounds for the true residual, the updated tolerance ζ satisfies

$$\zeta \geq \omega\frac{1-\omega}{1+\omega}\|\mathbf{f}-\mathbf{A}\mathbf{v}\|_{\ell^2} \geq \omega\frac{1-\omega}{1+\omega}(\eta-\zeta), \tag{2.5}$$

which implies $\zeta \geq \frac{\omega(1-\omega)}{1+2\omega-\omega^2}\eta$.

Let $\hat{\alpha}, \check{\alpha}, \lambda$ be available such that $\|\mathbf{A}\| \leq \hat{\alpha}$, $\|\mathbf{A}^{-1}\| \leq \check{\alpha}$ and $\|\mathbf{f}\|_{\ell^2} \leq \lambda$. Then $\kappa_{\mathbf{A}} := \hat{\alpha}\check{\alpha}$ is an upper bound for the condition number $\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ of $\mathbf{A}$.

The method $\texttt{Solve}_{\mathbf{A},\mathbf{f}}$ uses approximate residuals computed by $\texttt{Residual}_{\mathbf{A},\mathbf{f}}$ to adaptively select and iteratively solve a finite section of (2.1). For a finite $\Xi \subset \mathbb{N}$, a finitely supported $\mathbf{r} \in \ell^2$ and $\varepsilon > 0$, the routine

$$\texttt{Refine}[\Xi, \mathbf{r}, \varepsilon] \mapsto [\bar{\Xi}, \rho] \tag{2.6}$$

constructs a set $\bar{\Xi} \supset \Xi$ such that $\rho := \|\mathbf{r} - \mathbf{r}|_{\bar{\Xi}}\|_{\ell^2} \leq \varepsilon$, and $\#\bar{\Xi}$ is minimal with this property, up to a constant factor $\hat{c}$. This can be realized with $\hat{c} = 1$ by sorting $\mathbf{r}$ and appending the indices i to Ξ for which $|r_i|$ is largest. Using an approximate sorting routine, $\texttt{Refine}$ can be realized in linear complexity with respect to $\#\operatorname{supp}\mathbf{r}$ at the cost of a constant $\hat{c} > 1$.

$\mathtt{Solve}_{\mathbf{A},\mathbf{f}}[\varepsilon, \chi, \vartheta, \omega, \sigma, \beta] \mapsto [\mathbf{u}_\varepsilon, \bar{\varepsilon}]$

$\Xi^{(0)} \longleftarrow \varnothing$
$\tilde{\mathbf{u}}^{(0)} \longleftarrow \mathbf{0}$
$\delta_0 \longleftarrow \check{\alpha}^{1/2}\lambda$
for $k = 0, 1, 2, \ldots$ **do**
  if $\delta_k \leq \varepsilon$ **then break**
  $[\mathbf{r}_k, \eta_k, \zeta_k] \longleftarrow \mathtt{Residual}_{\mathbf{A},\mathbf{f}}[\varepsilon\check{\alpha}^{-1/2}, \tilde{\mathbf{u}}^{(k)}, \hat{\alpha}^{1/2}\delta_k, \chi, \omega, \beta]$
  $\bar{\delta}_k \longleftarrow \check{\alpha}^{1/2}(\eta_k + \zeta_k)$
  if $\bar{\delta}_k \leq \varepsilon$ **then break**
  $[\Xi^{(k+1)}, \rho_k] \longleftarrow \mathtt{Refine}[\Xi^{(k)}, \mathbf{r}_k, \sqrt{\eta_k^2 - (\zeta_k + \vartheta(\eta_k + \zeta_k))^2}]$
  $\bar{\vartheta}_k \longleftarrow \big(\sqrt{\eta_k^2 - \rho_k^2} - \zeta_k\big)/(\eta_k + \zeta_k)$
  $[\tilde{\mathbf{u}}^{(k+1)}, \tau_{k+1}] \longleftarrow \mathtt{Galerkin}_{\mathbf{A},\mathbf{f}}[\Xi^{(k+1)}, \tilde{\mathbf{u}}^{(k)}, \sigma \min(\delta_k, \bar{\delta}_k)]$
  $\delta_{k+1} \longleftarrow \tau_{k+1} + \sqrt{1 - \bar{\vartheta}_k^2 \kappa_{\mathbf{A}}^{-1}} \min(\delta_k, \bar{\delta}_k)$
$\mathbf{u}_\varepsilon \longleftarrow \tilde{\mathbf{u}}^{(k)}$
$\bar{\varepsilon} \longleftarrow \min(\delta_k, \bar{\delta}_k)$

The function

$$\mathtt{Galerkin}_{\mathbf{A},\mathbf{f}}[\Xi, \mathbf{v}, \varepsilon] \mapsto [\tilde{\mathbf{u}}, \tau] \tag{2.7}$$

approximates the solution of (2.1) restricted to the finite index set $\Xi \subset \mathbb{N}$ up to an error of at most $\tau \leq \varepsilon$ in the energy norm, using $\mathbf{v}$ as the initial approximation. For example, a conjugate gradient or conjugate residual method could be used to solve this linear system.

Remark 2.2 In the call of $\mathtt{Galerkin}_{\mathbf{A},\mathbf{f}}$ in $\mathtt{Solve}_{\mathbf{A},\mathbf{f}}$, the previous approximate solution is used as an initial approximation. Alternatively, the approximate residual $\mathbf{r}_k$, which is readily available, may be used to compute one step of a linear iteration, such as a Richardson method, prior to calling $\mathtt{Galerkin}_{\mathbf{A},\mathbf{f}}$. Although this may have quantitative advantages, we refrain from going into details in order to keep the presentation and analysis simple.

2.2 Convergence analysis

The convergence analysis of $\mathtt{Solve}_{\mathbf{A},\mathbf{f}}$ is based on [9, Lemma 4.1], which is the following lemma. We note that the solution of the restricted system (2.1) on a set $\Xi \subset \mathbb{N}$ is the Galerkin projection onto $\ell^2(\Xi) \subset \ell^2$.

Lemma 2.3 *Let $\Xi \subset \mathbb{N}$ and $\mathbf{v} \in \ell^2(\Xi)$ such that, for a $\vartheta \in [0, 1]$,*

$$\|(\mathbf{f} - \mathbf{A}\mathbf{v})|_\Xi\|_{\ell^2} \geq \vartheta \|\mathbf{f} - \mathbf{A}\mathbf{v}\|_{\ell^2}, \tag{2.8}$$

then the Galerkin projection $\bar{\mathbf{u}}$ of $\mathbf{u}$ onto $\ell^2(\Xi)$ satisfies

$$\|\mathbf{u} - \bar{\mathbf{u}}\|_{\mathbf{A}} \leq \sqrt{1 - \vartheta^2 \kappa_{\mathbf{A}}^{-1}} \|\mathbf{u} - \mathbf{v}\| \ . \tag{2.9}$$

We note that, by construction, if $\vartheta > 0$, $\omega > 0$ and $\omega + \vartheta + \omega\vartheta \leq 1$, then for all k, $\Xi^{(k+1)}$ in $\texttt{Solve}_{\mathbf{A},\mathbf{f}}$ is such that

$$\|(\mathbf{f} - \mathbf{A}\tilde{\mathbf{u}}^{(k)})|_{\Xi^{(k+1)}}\|_{\ell^2} \geq \bar{\vartheta}_k \|\mathbf{f} - \mathbf{A}\tilde{\mathbf{u}}^{(k)}\|_{\ell^2}\,, \tag{2.10}$$

and $\bar{\vartheta}_k \geq \vartheta$. Thus Lemma 2.3 implies an error reduction of at least $\sqrt{1 - \vartheta^2 \kappa_{\mathbf{A}}^{-1}}$ per step of $\texttt{Solve}_{\mathbf{A},\mathbf{f}}$, plus an error of τ_k in the approximation of the Galerkin projection.

Theorem 2.4 *If $\varepsilon > 0$, $\chi > 0$, $\vartheta > 0$, $\omega > 0$, $\omega + \vartheta + \omega\vartheta \leq 1$, $0 < \beta < 1$ and $0 < \sigma < 1 - \sqrt{1 - \vartheta^2 \kappa_{\mathbf{A}}^{-1}}$, then $\texttt{Solve}_{\mathbf{A},\mathbf{f}}[\varepsilon, \chi, \vartheta, \omega, \sigma, \beta]$ constructs a finitely supported $\mathbf{u}_\varepsilon$ with*

$$\|\mathbf{u} - \mathbf{u}_\varepsilon\|_{\mathbf{A}} \leq \bar{\varepsilon} \leq \varepsilon. \tag{2.11}$$

Moreover, for all $k \in \mathbb{N}_0$ reached by the iteration,

$$\kappa_{\mathbf{A}}^{-1/2} \frac{1 - \omega}{1 + \omega} \bar{\delta}_k \leq \|\mathbf{u} - \tilde{\mathbf{u}}^{(k)}\|_{\mathbf{A}} \leq \min(\delta_k, \bar{\delta}_k). \tag{2.12}$$

We refer to [27, Theorem 3.4] for a proof of Theorem 2.4, see also [22, Theorem 2.7].

Remark 2.5 Due to (2.12), in each call of $\texttt{Galerkin}_{\mathbf{A},\mathbf{f}}$, an error reduction of at most a fixed factor σ is required. Since the condition number of $\mathbf{A}$ restricted to any $\Xi \subset \mathbb{N}$ is at most $\kappa_{\mathbf{A}}$, a fixed number of steps of e.g. a conjugate gradient iteration suffice, even with no further preconditioning.

2.3 Optimality properties

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

$$\|\mathbf{v}\|_{\mathscr{A}^s} := \sup_{N \in \mathbb{N}_0} (N+1)^s \|\mathbf{v} - P_N(\mathbf{v})\|_{\ell^2} \tag{2.13}$$

and

$$\mathscr{A}^s := \left\{ \mathbf{v} \in \ell^2 \,;\; \|\mathbf{v}\|_{\mathscr{A}^s} < \infty \right\}. \tag{2.14}$$

Setting $\varepsilon = \|\mathbf{v} - P_N(\mathbf{v})\|_{\ell^2} - \eta$ with $\eta \geq 0$, it follows that

$$\|\mathbf{v}\|_{\mathscr{A}^s} = \sup_{\varepsilon > 0} \varepsilon \left(\min \left\{ N \in \mathbb{N}_0 \,;\; \|\mathbf{v} - P_N(\mathbf{v})\|_{\ell^2} \leq \varepsilon \right\} \right)^s, \tag{2.15}$$

which is used as the definition in [20]. If the index set $\mathbb{N}$ is replaced by a countable set Ξ, we will write $\mathscr{A}^s(\Xi)$ for $\mathscr{A}^s$.

By definition, the space $\mathscr{A}^s$ contains all $\mathbf{v} \in \ell^2$ that can be approximated by finitely supported vectors with a rate s,

$$\|\mathbf{v} - P_N(\mathbf{v})\|_{\ell^2} \leq \|\mathbf{v}\|_{\mathscr{A}^s}(N+1)^{-s} \qquad \forall N \in \mathbb{N}_0. \tag{2.16}$$

The following theorem states that this method recovers the optimal rate s whenever $\mathbf{u} \in \mathscr{A}^s$, i.e. the approximate Galerkin projections $\tilde{\mathbf{u}}^{(k)}$ converge to $\mathbf{u}$ at a rate of s with respect to $\#\Xi^{(k)}$, under some conditions on the parameters of $\texttt{Solve}_{\mathbf{A},\mathbf{f}}$.

Theorem 2.6 *If the conditions of Theorem 2.4 are fulfilled,*

$$\hat{\vartheta} := \frac{\vartheta(1+\omega)+2\omega}{1-\omega} < \kappa_{\mathbf{A}}^{-1/2}, \tag{2.17}$$

and $\mathbf{u} \in \mathscr{A}^s$ *for an* $s > 0$, *then for all* $k \in \mathbb{N}_0$ *reached by* $\texttt{Solve}_{\mathbf{A},\mathbf{f}}$,

$$\|\mathbf{u} - \tilde{\mathbf{u}}^{(k)}\|_{\ell^2} \leq 2^s \hat{c}^s \kappa_{\mathbf{A}} \tau^{-1} \rho (1-\rho^{1/s})^{-s} \frac{1+\omega}{1-\omega} \|\mathbf{u}\|_{\mathscr{A}^s} (\#\Xi^{(k)})^{-s} \tag{2.18}$$

with $\rho = \sigma + \sqrt{1-\vartheta^2 \kappa_{\mathbf{A}}^{-1}}$ *and* $\tau = \sqrt{1-\hat{\vartheta}^2 \kappa_{\mathbf{A}}}$.

The proof of Theorem 2.6 hinges on the following Lemma. We refer to [27, Theorem 4.2] and [20,22] for details. For a proof of Lemma 2.7, we refer to [27, Lemma 4.1]. See also [22, Lemma 2.1] and [20, Lemma 4.1].

Lemma 2.7 *Let* $\Xi^{(0)} \subset \mathbb{N}$ *be a finite set and* $\mathbf{v} \in \ell^2(\Xi^{(0)})$. *If* $0 < \hat{\vartheta} < \kappa_{\mathbf{A}}^{-1/2}$ *and* $\Xi^{(0)} \subset \Xi^{(1)} \subset \mathbb{N}$ *with*

$$\#\Xi^{(1)} \leq c \min\left\{\#\Xi \,;\; \Xi^{(0)} \subset \Xi,\, \|(\mathbf{f}-\mathbf{A}\mathbf{v})|_{\Xi}\|_{\ell^2} \geq \hat{\vartheta}\|\mathbf{f}-\mathbf{A}\mathbf{v}\|_{\ell^2}\right\} \tag{2.19}$$

for a $c \geq 1$, *then*

$$\#(\Xi^{(1)} \backslash \Xi^{(0)}) \leq c \min\left\{\#\hat{\Xi} \,;\; \hat{\Xi} \subset \mathbb{N},\, \|\mathbf{u}-\hat{\mathbf{u}}\|_{\mathbf{A}} \leq \tau\|\mathbf{u}-\mathbf{v}\|_{\mathbf{A}}\right\} \tag{2.20}$$

for $\tau = \sqrt{1-\hat{\vartheta}^2\kappa_{\mathbf{A}}^{1/2}}$, *where* $\hat{\mathbf{u}}$ *denotes the Galerkin projection of* $\mathbf{u}$ *onto* $\ell^2(\hat{\Xi})$.

Theorem 2.6 implies that the algorithm $\texttt{Solve}_{\mathbf{A},\mathbf{f}}$ is stable in $\mathscr{A}^s$. If the conditions of the theorem are satisfied, then for all k reached in the iteration,

$$\|\tilde{\mathbf{u}}^{(k)}\|_{\mathscr{A}^s} \leq \left(1 + \frac{2^{1+s}\hat{c}^s\kappa_{\mathbf{A}}\rho(1+\omega)}{\tau(1-\rho^{1/s})^s(1-\omega)}\right)\|\mathbf{u}\|_{\mathscr{A}^s}, \tag{2.21}$$

see e.g. [27, Lemma 4.6].

Remark 2.8 The sparsity of approximate solutions is of secondary importance compared to the computational cost of $\texttt{Solve}_{\mathbf{A},\mathbf{f}}$. Under suitable assumptions, the number of operations used by a call of $\texttt{Solve}_{\mathbf{A},\mathbf{f}}$ is on the order of $\varepsilon^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s}$, which is optimal due to (2.15). Besides the conditions of Theorem 2.6, this presumes that a call of $\texttt{Apply}_{\mathbf{A}}[\mathbf{v},\varepsilon]$ has a computational cost on the order of

$$1+\#\operatorname{supp}\mathbf{v}+\varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}, \tag{2.22}$$

and similarly the cost of $\texttt{RHS}_{\mathbf{f}}[\varepsilon]$ is $\mathcal{O}(\varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s})$. Due to the geometric decrease of the tolerances ζ in $\texttt{Residual}_{\mathbf{A},\mathbf{f}}$, the total cost of this routine is equivalent to that of the last iteration, which is $\mathcal{O}(\zeta_k^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s})$, using Theorem 2.6 and (2.21). This includes the cost of $\texttt{Refine}$ if this is realized by an approximate sorting routine with linear complexity. Finally, since only a fixed number of steps of a linear iteration is required in $\texttt{Galerkin}_{\mathbf{A},\mathbf{f}}$ by Remark 2.5, and each step can realistically be performed in at most the same complexity as $\texttt{Apply}_{\mathbf{A}}$, the computational cost of the k-th iteration in $\texttt{Solve}_{\mathbf{A},\mathbf{f}}$ is $\mathcal{O}(\zeta_k^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s})$. Equation (2.5) implies that this is equivalent to $\mathcal{O}(\bar{\delta}_k^{-1/s}\|\mathbf{u}\|_{\mathscr{A}^s}^{1/s})$, and since the error estimates $\bar{\delta}_k$ decrease geometrically, the total cost of $\texttt{Solve}_{\mathbf{A},\mathbf{f}}$ is dominated by that of the last iteration of the loop, in which the error is on the order of ε.

3 Greedy algorithms

3.1 A generalized knapsack problem

We consider a discrete optimization problem in which both the objective and the constraints are given by sums over an arbitrary set $\mathscr{M}\subset\mathbb{N}_0$. For each $m\in\mathscr{M}$, we have two increasing sequences $(c_j^m)_{j\in\mathbb{N}_0}$ and $(\omega_j^m)_{j\in\mathbb{N}_0}$ defining costs and values: for any integer sequence $\mathbf{j}=(j_m)_{m\in\mathscr{M}}\in\mathbb{N}_0^{\mathscr{M}}$, the total cost of $\mathbf{j}$ is

$$c_{\mathbf{j}}:=\sum_{m\in\mathscr{M}}c_{j_m}^m \tag{3.1}$$

and the total value of $\mathbf{j}$ is

$$\omega_{\mathbf{j}}:=\sum_{m\in\mathscr{M}}\omega_{j_m}^m. \tag{3.2}$$

Our goal is to maximize $\omega_{\mathbf{j}}$ under a constraint on $c_{\mathbf{j}}$, or to minimize $c_{\mathbf{j}}$ under a constraint on $\omega_{\mathbf{j}}$. We consider $\mathbf{j}\in\mathbb{N}_0^{\mathscr{M}}$ *optimal* if $c_{\mathbf{i}}\le c_{\mathbf{j}}$ implies $\omega_{\mathbf{i}}\le\omega_{\mathbf{j}}$ or, equivalently, $\omega_{\mathbf{i}}>\omega_{\mathbf{j}}$ implies $c_{\mathbf{i}}>c_{\mathbf{j}}$.

Remark 3.1 The classical knapsack problem is equivalent to the above optimization problem in the case that $\mathscr{M}$ is finite, and for all $m\in\mathscr{M}$, $\omega_0^m=0$ and $\omega_j^m=\omega_1^m$ for all $j\ge 1$. Then without loss of generality, we can set $c_0^m:=0$ for all $m\in\mathscr{M}$,

and the values c_j^m for $j \geq 2$ are irrelevant due to the assumption that $(c_j^m)_{j\in\mathbb{N}_0}$ is increasing. Optimal sequences $\mathbf{j} \in \mathbb{N}_0^{\mathscr{M}}$ will only take the values 0 and 1, and can thus be interpreted as subsets of $\mathscr{M}$.

We note that greedy methods only construct a sequence of optimal solutions. They do not maximize $\omega_{\mathbf{j}}$ under an arbitrary constraint on $c_{\mathbf{j}}$, and thus do not solve an NP-hard problem.

Remark 3.2 We are particularly interested in minimizing an error under constraints on the computational cost of an approximation with this error tolerance. Given sequences $(e_j^m)_{j\in\mathbb{N}_0}$ and $(c_j^m)_{j\in\mathbb{N}_0}$ of errors and corresponding costs, we define a sequence of values by $\omega_j^m := -e_j^m$. If $(e_j^m)_{j\in\mathbb{N}_0}$ is decreasing, then $(\omega_j^m)_{j\in\mathbb{N}_0}$ is increasing. Typically, as $j \to \infty$, we have $e_j^m \to 0$ and $c_j^m \to \infty$. Then, although it is increasing, $(\omega_j^m)_{j\in\mathbb{N}_0}$ remains bounded, and it is reasonable to assume that $(\omega_j^m)_{j\in\mathbb{N}_0}$ increases more slowly than $(c_j^m)_{j\in\mathbb{N}_0}$, in a sense that is made precise below.

3.2 A sequence of optimal solutions

We iteratively construct a sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$ in $\mathbb{N}_0^{\mathscr{M}}$ such that, under some assumptions, each $\mathbf{j}^k = (j_m^k)_{m\in\mathscr{M}}$ is optimal. For all $m \in \mathscr{M}$ and all $j \in \mathbb{N}_0$, let

$$\Delta c_j^m := c_{j+1}^m - c_j^m \qquad \text{and} \qquad \Delta\omega_j^m := \omega_{j+1}^m - \omega_j^m. \tag{3.3}$$

Furthermore, let q_j^m denote the quotient of these two increments,

$$q_j^m := \frac{\Delta\omega_j^m}{\Delta c_j^m}, \quad j \in \mathbb{N}_0, \tag{3.4}$$

which can be interpreted as the value to cost ratio of passing from j to $j+1$ in the index $m \in \mathscr{M}$.

Let $\mathbf{j}^0 := \mathbf{0} \in \mathbb{N}_0^{\mathscr{M}}$. For all $k \in \mathbb{N}_0$, we construct $\mathbf{j}^{k+1}$ from $\mathbf{j}^k$ as follows. Let $m_k = m \in \mathbb{N}_0$ maximize $q_{j_m^k}^m$. Existence of such maxima is ensured by the last statement in Assumption 3.A. If the maximum is not unique, select m_k to be minimal among all maxima. Then define $j_{m_k}^{k+1} := j_{m_k}^k + 1$, and set $j_m^{k+1} := j_m^k$ for all $m \neq m_k$. For this sequence, we abbreviate $c_k := c_{\mathbf{j}^k}$ and $\omega_k := \omega_{\mathbf{j}^k}$.

Assumption 3.A For all $m \in \mathscr{M}$,

$$c_0^m = 0 \qquad \text{and} \quad \Delta c_j^m > 0 \qquad \forall j \in \mathbb{N}_0, \tag{3.5}$$

i.e. $(c_j^m)_{j\in\mathbb{N}_0}$ is strictly increasing. Also, $(\omega_0^m)_{m\in\mathscr{M}} \in \ell^1(\mathscr{M})$ and $(\omega_j^m)_{j\in\mathbb{N}_0}$ is nondecreasing for all $m \in \mathscr{M}$, i.e. $\Delta\omega_j^m \geq 0$ for all $j \in \mathbb{N}_0$. Furthermore, for each $m \in \mathscr{M}$, the sequence $(q_j^m)_{j\in\mathbb{N}_0}$ is nonincreasing, i.e. if $i \geq j$, then $q_i^m \leq q_j^m$. Finally, for any $\varepsilon > 0$, there are only finitely many $m \in \mathscr{M}$ for which $q_0^m \geq \varepsilon$.

The assumption that $(q_j^m)_{j\in\mathbb{N}_0}$ is nonincreasing is equivalent to

$$\frac{\Delta\omega_i^m}{\Delta\omega_j^m} \leq \frac{\Delta c_i^m}{\Delta c_j^m} \quad \text{if} \quad i \geq j \tag{3.6}$$

if $\Delta\omega_j^m > 0$. In this sense, $(\omega_j^m)_{j\in\mathbb{N}_0}$ increases more slowly than $(c_j^m)_{j\in\mathbb{N}_0}$. Also, this assumption implies that if $\Delta\omega_j^m = 0$, then $\omega_i^m = \omega_j^m$ for all $i \geq j$.

We define a total order on $\mathscr{M} \times \mathbb{N}_0$ by

$$(m, j) \prec (n, i) \quad \text{if} \quad \begin{cases} q_j^m > q_i^n & \text{or} \\ q_j^m = q_i^n & \text{and} \quad m < n \quad \text{or} \\ q_j^m = q_i^n & \text{and} \quad m = n \quad \text{and} \ j < i\ . \end{cases} \tag{3.7}$$

To any sequence $\mathbf{j} = (j_m)_{m\in\mathscr{M}}$ in $\mathbb{N}_0$, we associate the set

$$\{\{\mathbf{j}\}\} := \{(m, j) \in \mathscr{M} \times \mathbb{N}_0\ ;\ j < j_m\}\,. \tag{3.8}$$

Lemma 3.3 *For all $k \in \mathbb{N}_0$, $\{\{k\}\} := \{\{\mathbf{j}^k\}\}$ consists of the first k terms of $\mathscr{M} \times \mathbb{N}_0$ with respect to the order $\prec$.*

Proof The assertion is trivial for $k = 0$. Assume it holds for some $k \in \mathbb{N}_0$. By definition,

$$\{\{k+1\}\} = \{\{k\}\} \cup \{(m_k, j_{m_k}^k)\},$$

and $(m_k, j_{m_k}^k)$ is the $\prec$-minimal element of the set $\{(m, j_m^k)\ ;\ m \in \mathscr{M}\}$. For each $m \in \mathscr{M}$, Assumption 3.A implies $q_i^m \leq q_{j_m^k}^m$ for all $i \geq j_m^k + 1$. Therefore, $(m, j_m^k) \prec (m, i)$ for all $i \geq j_m^k + 1$, and consequently $(m_k, j_{m_k}^k)$ is the $\prec$-minimal element of $(\mathscr{M} \times \mathbb{N}_0)\backslash\{\{k\}\}$.

Theorem 3.4 *For all $k \in \mathbb{N}_0$, the sequence $\mathbf{j}^k$ maximizes $\omega_{\mathbf{j}}$ among all finitely supported sequences $\mathbf{j} = (j_m)_{m\in\mathscr{M}}$ in $\mathbb{N}_0$ with $c_{\mathbf{j}} \leq c_k$. Furthermore, if $c_{\mathbf{j}} < c_k$ and there exist k pairs $(m, i) \in \mathscr{M} \times \mathbb{N}_0$ with $\Delta\omega_i^m > 0$, then $\omega_{\mathbf{j}} < \omega_k$.*

Proof Let $k \in \mathbb{N}$ and let $\mathbf{j} = (j_m)_{m\in\mathscr{M}}$ be a finitely supported sequence in $\mathbb{N}_0$ with $c_{\mathbf{j}} \leq c_k$. By definition,

$$\omega_{\mathbf{j}} = \sum_{m\in\mathscr{M}} \omega_0^m + \sum_{m\in\mathscr{M}} \sum_{i=0}^{j_m-1} q_i^m \Delta c_i^m = \omega_{\mathbf{j}^0} + \sum_{(m,i)\in\{\{\mathbf{j}\}\}} q_i^m \Delta c_i^m.$$

Therefore, the assertion reduces to

$$\sum_{(m,i)\in\{\{\mathbf{j}\}\}\backslash\{\{k\}\}} q_i^m \Delta c_i^m \leq \sum_{(m,i)\in\{\{k\}\}\backslash\{\{\mathbf{j}\}\}} q_i^m \Delta c_i^m.$$

Note that by (3.1) and (3.3),

$$\sum_{(m,i)\in\{\{\mathbf{j}\}\}\setminus\{\{k\}\}} \Delta c_i^m = c_{\mathbf{j}} - c' \quad \text{for} \quad c' := \sum_{(m,i)\in\{\{\mathbf{j}\}\}\cap\{\{k\}\}} \Delta c_i^m.$$

By Lemma 3.3 and (3.7), $q := q_{j_{m_{k-1}}^{k-1}}^{m_{k-1}}$ satisfies $q \le q_i^m$ for all $(m,i) \in \{\{k\}\}$, and $q_i^m \le q$ for all $(m,i) \in (\mathscr{M}\times\mathbb{N}_0)\setminus\{\{k\}\}$. In particular, $q > 0$ if there exist k pairs $(m,i) \in \mathscr{M}\times\mathbb{N}_0$ with $q_i^m > 0$ since $\#\{\{k\}\} = k$. Consequently,

$$\begin{aligned}\sum_{(m,i)\in\{\{\mathbf{j}\}\}\setminus\{\{k\}\}} q_i^m \Delta c_i^m &\le q \sum_{(m,i)\in\{\{\mathbf{j}\}\}\setminus\{\{k\}\}} \Delta c_i^m = q(c_{\mathbf{j}} - c')\\ &\le q(c_k - c') \le \sum_{(m,i)\in\{\{k\}\}\setminus\{\{\mathbf{j}\}\}} q_i^m \Delta c_i^m,\end{aligned}$$

and this inequality is strict if $q > 0$ and $c_k > c_{\mathbf{j}}$.

Similarly, $\mathbf{j}^k$ also minimizes $c_{\mathbf{j}}$ among $\mathbf{j}$ with $\omega_{\mathbf{j}} \ge \omega_k$.

3.3 Numerical construction

We consider numerical methods for constructing the sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$ from Sect. 3.2. To this end, we assume that, for each $m \in \mathscr{M}$, the sequences $(c_j^m)_{j\in\mathbb{N}_0}$ and $(\omega_j^m)_{j\in\mathbb{N}_0}$ are stored as linked lists.

Initially, we consider the case that $\mathscr{M}$ is finite with $\#\mathscr{M} =: M$. To construct $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$, we use a list $\mathscr{N}$ of the triples $(m, j_m^k, q_{j_m^k}^m)$, sorted in ascending order with respect to $\prec$. This list may be realized as a linked list or as a tree. The data structure must provide functions `PopMin` for removing the minimal element from the list, and `Insert` for inserting a new element into the list.

`NextOpt`$[\mathbf{j}, \mathscr{N}] \mapsto [\mathbf{j}, m, \mathscr{N}]$

$m \longleftarrow$ `PopMin`$(\mathscr{N})$
$j_m \longleftarrow j_m + 1$
$q \longleftarrow (\omega_{j_m+1}^m - \omega_{j_m}^m)/(c_{j_m+1}^m - c_{j_m}^m)$
$\mathscr{N} \longleftarrow$ `Insert`$(\mathscr{N}, (m, j_m, q))$

Proposition 3.5 *Let $\mathscr{N}_0$ be initialized as $\{(m,0,q_0^m)\,;\ m\in\mathscr{M}\}$ and $\mathbf{j}^0 := \mathbf{0} \in \mathbb{N}_0^{\mathscr{M}}$. Then the recursive application of*

$$\texttt{NextOpt}[\mathbf{j}^k, \mathscr{N}_k] \mapsto [\mathbf{j}^{k+1}, m_k, \mathscr{N}_{k+1}] \tag{3.9}$$

constructs the sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$ as defined above. Initialization of the data structure $\mathscr{N}_0$ requires $\mathscr{O}(M)$ memory and $\mathscr{O}(M\log M)$ operations. One step of (3.9) requires $\mathscr{O}(M)$ operations if $\mathscr{N}$ is realized as a linked list, and $\mathscr{O}(\log M)$ operations if $\mathscr{N}$ is realized as a tree. The total number of operations required by the first k steps is $\mathscr{O}(kM)$ in the former case and $\mathscr{O}(k\log M)$ in the latter. In both cases, the total memory requirement for the first k steps is $\mathscr{O}(M+k)$.

Proof Recursive application of `NextOpt` as in (3.9) constructs the sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$ by Lemma 3.3 and the definition of $\prec$. In the k-th step, the element m_k is removed from $\mathcal{N}$ and reinserted in a new position. Therefore, the size of $\mathcal{N}$ remains constant at M. The computational cost of (3.9) is dominated by the insert operation on $\mathcal{N}$, which has the complexity stated above.

We turn to the case that $\mathcal{M}$ is countably infinite. By enumerating the elements of $\mathcal{M}$, it suffices to consider $\mathcal{M} = \mathbb{N}$. We assume in this case that the sequence $(q_0^m)_{m\in\mathcal{M}}$ is nonincreasing.

As above, we use a list $\mathcal{N}$ of triples $(m, j_m^k, q_{j_m^k}^m)$ to construct the sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$. However, $\mathcal{N}$ should only store triples for which m is a candidate for the next value of m_k, i.e. all m with $j_m^k \neq 0$ and the smallest m with $j_m^k = 0$. As in the finite case, $\mathcal{N}$ can be realized as a linked list or a tree. The data structure should provide functions `PopMin` for removing the smallest element with respect to the ordering $\prec$, and `Insert` for inserting a new element.

`NextOptInf`$[\mathbf{j}, \mathcal{N}, M] \mapsto [\mathbf{j}, m, \mathcal{N}, M]$

$m \longleftarrow$ `PopMin`$(\mathcal{N})$
$j_m \longleftarrow j_m + 1$
$q \longleftarrow (\omega_{j_m+1}^m - \omega_{j_m}^m)/(c_{j_m+1}^m - c_{j_m}^m)$
$\mathcal{N} \longleftarrow$ `Insert`$(\mathcal{N}, (m, j_m, q))$
if $m = M$ **then**
  $M \longleftarrow M + 1$
  $q \longleftarrow (\omega_1^M - \omega_0^M)/c_1^M$
  $\mathcal{N} \longleftarrow$ `Insert`$(\mathcal{N}, (M, 1, q))$

Proposition 3.6 *Let $\mathcal{N}_0$ be initialized as $\{(1, 0, q^1)\}$, $M_0 := 1$ and $\mathbf{j}^0 := \mathbf{0} \in \mathbb{N}_0^{\mathcal{M}}$. Then the recursion*

$$\texttt{NextOptInf}[\mathbf{j}^k, \mathcal{N}_k, M_k] \mapsto [\mathbf{j}^{k+1}, m_k, \mathcal{N}_{k+1}, M_{k+1}] \tag{3.10}$$

constructs the sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$ as defined above. For all $k \in \mathbb{N}_0$, the ordered set $\mathcal{N}_k$ contains exactly M_k elements, and $M_k \leq k$. The k-th step of (3.10) requires $\mathcal{O}(k)$ operations if $\mathcal{N}$ is realized as a linked list, and $\mathcal{O}(\log k)$ operations if $\mathcal{N}$ is realized as a tree. The total number of operations required by the first k steps is $\mathcal{O}(k^2)$ in the former case and $\mathcal{O}(k \log k)$ in the latter. In both cases, the total memory requirement for the first k steps is $\mathcal{O}(k)$.

Proof It follows from the definitions that recursive application of `NextOptInf` as in (3.10) constructs the sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$. In the k-th step, the element m_k is removed from $\mathcal{N}$ and reinserted in a new position. If $m_k = M$, an additional element is inserted, and M is incremented. Therefore, the number of elements in $\mathcal{N}$ is M, and $M \leq k$. The computational cost of (3.10) is dominated by the insert operation on $\mathcal{N}$, which has the complexity stated above, see e.g. [13].

Remark 3.7 As mentioned above, $(c_j^m)_{j\in\mathbb{N}_0}$ and $(\omega_j^m)_{j\in\mathbb{N}_0}$ are assumed to be stored in a linked list for each $m \in \mathscr{M}$. By removing the first element from the $\mathscr{M}_k$-th list in the k-th step of (3.9) or (3.10), `NextOpt` and `NextOptInf` only ever access the first two elements of one of these lists, which takes $\mathscr{O}(1)$ time. The memory locations of the lists can be stored in a hash table for efficient access.

Remark 3.8 An appropriate way to store $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$ is to collect $(m_k)_{k\in\mathbb{N}_0}$ in a linked list. Then $\mathbf{j}^k$ can be reconstructed by reading the first k elements of this list, which takes $\mathscr{O}(k)$ time independently of the size of the list. Also, the total memory requirement is $\mathscr{O}(\bar{k})$ if the first $\bar{k}$ elements are stored.

4 Adaptive application of s^*-compressible operators

4.1 s^*-compressibility and s^*-computability

A routine $\texttt{Apply}_{\mathbf{A}}$ for approximately applying an operator $\mathbf{A} \in \mathscr{L}(\ell^2)$ to a finitely supported vector constitutes an essential component of the adaptive solver from Sect. 2. Such a routine can be constructed if $\mathbf{A}$ can be approximated by sparse operators, as in the following definition. Again, we interpret $\mathbf{A} \in \mathscr{L}(\ell^2)$ also as a bi-infinite matrix, and restrict to the index set $\mathbb{N}$ only to simplify notation.

Definition 4.1 An operator $\mathbf{A} \in \mathscr{L}(\ell^2)$ is *n-sparse* if each column contains at most n nonzero entries. It is *s^*-compressible* for an $s^* \in (0, \infty]$ if there exists a sequence $(\mathbf{A}_j)_{j\in\mathbb{N}}$ in $\mathscr{L}(\ell^2)$ such that $\mathbf{A}_j$ is n_j-sparse with $(n_j)_{j\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}}$ satisfying

$$c_{\mathbf{A}} := \sup_{j\in\mathbb{N}} \frac{n_{j+1}}{n_j} < \infty \tag{4.1}$$

and for every $s \in (0, s^*)$,

$$d_{\mathbf{A},s} := \sup_{j\in\mathbb{N}} n_j^s \|\mathbf{A} - \mathbf{A}_j\|_{\ell^2\to\ell^2} < \infty. \tag{4.2}$$

The operator $\mathbf{A}$ is *strictly s^*-compressible* if, in addition,

$$\sup_{s\in(0,s^*)} d_{\mathbf{A},s} < \infty. \tag{4.3}$$

Remark 4.2 Equation (4.2) states that for all $s \in (0, s^*)$, the approximation errors satisfy

$$e_{\mathbf{A},j} := \|\mathbf{A} - \mathbf{A}_j\|_{\ell^2\to\ell^2} \le d_{\mathbf{A},s} n_j^{-s}, \quad j \in \mathbb{N}. \tag{4.4}$$

If $s^* < \infty$, this is equivalent to the condition that $(n_j^{s^*} e_{\mathbf{A},j})_{j\in\mathbb{N}}$ grows subalgebraically in n_j, i.e.

$$n_j^{s^*} e_{\mathbf{A},j} \le \inf_{r>0} d_{\mathbf{A},s^*-r} n_j^r, \quad j \in \mathbb{N}. \tag{4.5}$$

Strict s^*-compressibility states that the right hand side of (4.5) is bounded in j, i.e.

$$d_{\mathbf{A},s^*} = \sup_{j\in\mathbb{N}} n_j^{s^*} e_{\mathbf{A},j} = \sup_{j\in\mathbb{N}} \sup_{s\in(0,s^*)} n_j^s e_{\mathbf{A},j} = \sup_{s\in(0,s^*)} d_{\mathbf{A},s} < \infty. \tag{4.6}$$

Of course, s^*-compressibility implies strict s-compressibility for all $s \in (0, s^*)$.

Proposition 4.3 *Let $\mathbf{A} \in \mathscr{L}(\ell^2)$ be s^*-compressible with an approximating sequence $(\mathbf{A}_j)_{j\in\mathbb{N}}$ as in Definition 4.1, and set $\mathbf{A}_0 := \mathbf{0}$. There is a map $j : [0,\infty) \to \mathbb{N}_0$ such that $\mathbf{A}_{j(r)}$ is r-sparse for all $r \in [0,\infty)$ and for all $s \in (0, s^*)$,*

$$e_{\mathbf{A},j(r)} = \|\mathbf{A} - \mathbf{A}_{j(r)}\|_{\ell^2\to\ell^2} \le \max\left(c_{\mathbf{A}}^s d_{\mathbf{A},s}, n_1^s e_{\mathbf{A},0}\right) r^{-s} \tag{4.7}$$

for $r > 0$, where $e_{\mathbf{A},0} := \|\mathbf{A}\|_{\ell^2\to\ell^2}$.

Proof Set $n_0 := 0$ and define

$$j(r) := \max\left\{ j \in \mathbb{N}_0 \,;\, n_j \le r \right\}, \quad r \in [0,\infty). \tag{4.8}$$

Then $\mathbf{A}_{j(r)}$ is r-sparse, and if $j(r) \ge 1$,

$$e_{\mathbf{A},j(r)} \le d_{\mathbf{A},s} n_{j(r)}^{-s} \le d_{\mathbf{A},s} c_{\mathbf{A}}^s n_{j(r)+1}^{-s} \le d_{\mathbf{A},s} c_{\mathbf{A}}^s r^{-s}$$

by (4.4) and (4.1). If $j(r) = 0$, then $r < n_1$, and

$$e_{\mathbf{A},j(r)} = e_{\mathbf{A},0} \le e_{\mathbf{A},0} n_1^s r^{-s}.$$

In particular, Proposition 4.3 implies that Definition 4.1 coincides with the notion of s^*-compressibility for example in [22,32], i.e. one can assume $n_j = j$ in the definition of s^*-compressibility at the cost of increasing the constants (4.2) and obscuring the discrete structure of the sparse approximating sequence. We denote the resulting compressibility constants by

$$\tilde{d}_{\mathbf{A},s} := \sup_{r\in(0,\infty)} r^s \|\mathbf{A} - \mathbf{A}_{j(r)}\|_{\ell^2\to\ell^2} \le \max\left(c_{\mathbf{A}}^s d_{\mathbf{A},s}, n_1^s e_{\mathbf{A},0}\right) < \infty \tag{4.9}$$

for $s \in (0, s^*)$, where $j(r)$ is given by (4.8). Also, it follows using Proposition 4.3 that any symmetric s^*-compressible operator $\mathbf{A}$ is in the class $\mathscr{B}_s$, as defined in [9], for all $s \in [0, s^*)$.

Although s^*-compressibility is a precise mathematical property, it is only useful for applications if the sparse approximations to the bi-infinite matrix can be computed efficiently. This is the context of the following, more restrictive definition.

Definition 4.4 An operator $\mathbf{A} \in \mathscr{L}(\ell^2)$ is *s^*-computable* for an $s^* \in (0,\infty]$ if it is s^*-compressible with an approximating sequence $(\mathbf{A}_j)_{j\in\mathbb{N}}$ as in Definition 4.1 such that $\mathbf{A}_j$ is n_j-sparse and there exists a routine

$$\texttt{Build}_{\mathbf{A}}[j,k] \mapsto \left[(l_i)_{i=1}^{n_j}, (a_i)_{i=1}^{n_j} \right] \tag{4.10}$$

such that the k-th column of $\mathbf{A}_j$ is equal to $\sum_{i=1}^{n_j} a_i \varepsilon_{l_i}$, where ε_{l_i} is the Kronecker sequence that is 1 at l_i and 0 elsewhere, and there is a constant $b_{\mathbf{A}}$ such that the number of arithmetic operations and storage locations used by a call of $\texttt{Build}_{\mathbf{A}}[j,k]$ is less than $b_{\mathbf{A}} n_j$ for any $j \in \mathbb{N}$ and $k \in \mathbb{N}$.

Note that the indices l_i in (4.10) are not assumed to be distinct, so a single entry of $\mathbf{A}_j$ may be given by a sum of values a_i. However, the total number of a_i computed by $\texttt{Build}_{\mathbf{A}}[j,k]$ is at most n_j.

4.2 An adaptive approximate multiplication routine

It was shown in [9,10] that s^*-computable operators can be applied efficiently to finitely supported vectors. A routine with computational advantages was presented in [20]. We extend this method by using a greedy algorithm to solve the optimization problem at the heart of the routine.

Let $\mathbf{A} \in \mathscr{L}(\ell^2)$ and for all $k \in \mathbb{N}_0$, let $\mathbf{A}_k$ be n_k-sparse with $n_0 = 0$ and

$$\|\mathbf{A} - \mathbf{A}_k\|_{\ell^2 \to \ell^2} \le \bar{e}_{\mathbf{A},k}. \tag{4.11}$$

We consider a partitioning of a vector $\mathbf{v} \in \ell^2$ into $\mathbf{v}_{[p]} := \mathbf{v}|_{\Xi_p}$, $p = 1, \dots, P$, for disjoint index sets $\Xi_p \subset \mathbb{N}$. This can be approximate in that $\mathbf{v}_{[1]} + \cdots + \mathbf{v}_{[P]}$ only approximates $\mathbf{v}$ in ℓ^2. We think of $\mathbf{v}_{[1]}$ as containing the largest elements of $\mathbf{v}$, $\mathbf{v}_{[2]}$ the next largest, and so on.

Such a partitioning can be constructed by the approximate sorting algorithm

$$\texttt{BucketSort}[\mathbf{v}, \varepsilon] \mapsto \left[(\mathbf{v}_{[p]})_{p=1}^{P}, (\Xi_p)_{p=1}^{P} \right], \tag{4.12}$$

which, given a finitely supported $\mathbf{v} \in \ell^2$ and a threshold $\varepsilon > 0$, returns index sets

$$\Xi_p := \left\{ \mu \in \mathbb{N} \,;\, |v_\mu| \in (2^{-p/2}\|\mathbf{v}\|_{\ell^\infty}, 2^{-(p-1)/2}\|\mathbf{v}\|_{\ell^\infty}] \right\} \tag{4.13}$$

and $\mathbf{v}_{[p]} := \mathbf{v}|_{\Xi_p}$, see [2,20,22,29]. The integer P is minimal with

$$2^{-P/2}\|\mathbf{v}\|_{\ell^\infty}\sqrt{\#\operatorname{supp}\mathbf{v}} \le \varepsilon. \tag{4.14}$$

By [22, Rem. 2.3] or [20, Prop. 4.4], the number of operations and storage locations required by a call of $\texttt{BucketSort}[\mathbf{v}, \varepsilon]$ is bounded by

$$\#\operatorname{supp}\mathbf{v} + \max(1, \lceil \log(\|\mathbf{v}\|_{\ell^\infty}\sqrt{\#\operatorname{supp}\mathbf{v}}/\varepsilon) \rceil). \tag{4.15}$$

For any $\mathbf{k} = (k_p)_{p=1}^{\ell} \in \mathbb{N}_0^\ell$, with $\ell \in \mathbb{N}_0$ determined as in $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$, let

$$\zeta_{\mathbf{k}} := \sum_{p=1}^{\ell} \bar{e}_{\mathbf{A},k_p} \|\mathbf{v}_{[p]}\|_{\ell^2(\Xi_p)} \quad \text{and} \quad \sigma_{\mathbf{k}} := \sum_{p=1}^{\ell} n_{k_p}(\#\operatorname{supp}\mathbf{v}_{[p]}). \tag{4.16}$$

$\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon] \mapsto \mathbf{z}$

$(\mathbf{v}_{[p]})_{p=1}^{P} \longleftarrow \texttt{BucketSort}\left[\mathbf{v}, \dfrac{\varepsilon}{2\bar{e}_{\mathbf{A},0}}\right]$

compute the minimal $\ell \in \{0, 1, \ldots, P\}$ s.t. $\delta := \bar{e}_{\mathbf{A},0}\|\mathbf{v} - \sum_{p=1}^{\ell} \mathbf{v}_{[p]}\|_{\ell^2} \leq \dfrac{\varepsilon}{2}$

$\mathbf{k} = (k_p)_{p=1}^{\ell} \longleftarrow (0)_{p=1}^{\ell}$

while $\zeta_{\mathbf{k}} > \varepsilon - \delta$ **do**

  $\mathbf{k} \longleftarrow \texttt{NextOpt}[\mathbf{k}]$ with objective $-\zeta_{\mathbf{k}}$ and cost $\sigma_{\mathbf{k}}$

$\mathbf{z} \longleftarrow \sum_{p=1}^{\ell} \mathbf{A}_{k_p} \mathbf{v}_{[p]}$

The algorithm $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ has three distinct parts. First, the elements of $\mathbf{v}$ are grouped according to their magnitude. Elements smaller than a certain tolerance are neglected. This truncation of the vector $\mathbf{v}$ produces an error of at most $\delta \leq \varepsilon/2$.

Next, the greedy algorithm from Sect. 3 is used to assign to each segment $\mathbf{v}_{[p]}$ of $\mathbf{v}$ a sparse approximation $\mathbf{A}_{k_p}$ of $\mathbf{A}$. Starting with $\mathbf{A}_{k_p} = \mathbf{0}$ for all $p = 1, \ldots, \ell$, these approximations are refined iteratively until an estimate for the error resulting from the approximation of $\mathbf{A}$ by $\mathbf{A}_{k_p}$ for all $p = 1, \ldots, \ell$ is bounded by $\zeta_{\mathbf{k}} \leq \varepsilon - \delta$.

Finally, the multiplications determined by the previous two steps are performed. A few elementary properties of this method are summarized in the following proposition.

Proposition 4.5 *For any finitely supported $\mathbf{v} \in \ell^2$ and any $\varepsilon > 0$, if $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ terminates, its output is a finitely supported $\mathbf{z} \in \ell^2$ with*

$$\#\operatorname{supp} \mathbf{z} \leq \sum_{p=1}^{\ell} n_{k_p} (\#\operatorname{supp} \mathbf{v}_{[p]}) \tag{4.17}$$

and

$$\|\mathbf{A}\mathbf{v} - \mathbf{z}\|_{\ell^2} \leq \delta + \zeta_{\mathbf{k}} \leq \varepsilon, \tag{4.18}$$

where $\mathbf{k} = (k_p)_{p=1}^{\ell}$ is the vector constructed by the greedy algorithm in $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$. Furthermore, the number of arithmetic operations in the final step of $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ is bounded by

$$\sum_{p=1}^{\ell} n_{k_p} (\#\operatorname{supp} \mathbf{v}_{[p]}) \tag{4.19}$$

if the relevant entries of $\mathbf{A}_{k_p}$ are precomputed.

Proof We show (4.18). Since $\|\mathbf{A}\|_{\ell^2 \to \ell^2} \leq \bar{e}_{\mathbf{A},0}$,

$$\left\| \mathbf{A}\mathbf{v} - \mathbf{A} \sum_{p=1}^{\ell} \mathbf{v}_{[p]} \right\| \leq \bar{e}_{\mathbf{A},0} \left\| \mathbf{v} - \sum_{p=1}^{\ell} \mathbf{v}_{[p]} \right\| = \delta \leq \frac{\varepsilon}{2}.$$

By (4.11), if $\mathbf{k} = (k_p)_{p=1}^{\ell}$ is the final value of $\mathbf{k}$,

$$\sum_{p=1}^{\ell} \|\mathbf{A}\mathbf{v}_{[p]} - \mathbf{A}_{k_p}\mathbf{v}_{[p]}\|_{\ell^2} \le \sum_{p=1}^{\ell} \bar{e}_{\mathbf{A},k_p}\|\mathbf{v}_{[p]}\|_{\ell^2(\Xi_p)} = \zeta_{\mathbf{k}} \le \varepsilon - \delta.$$

Let $\mathbf{v} \in \ell^2$ be finitely supported and $\varepsilon > 0$. Note that by (4.13) and (4.14),

$$\left\| \mathbf{v} - \sum_{p=1}^{P} \mathbf{v}_{[p]} \right\| \le 2^{-P/2}\|\mathbf{v}\|_{\ell^\infty}\sqrt{\#\operatorname{supp}\mathbf{v}} \le \frac{\varepsilon}{2\bar{e}_{\mathbf{A},0}},$$

so ℓ is well-defined. It is not immediately clear, however, that the greedy algorithm in $\mathtt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ terminates. For all $k \in \mathbb{N}_0$, let

$$\eta_k := \frac{\bar{e}_{\mathbf{A},k} - \bar{e}_{\mathbf{A},k+1}}{n_{k+1} - n_k}. \tag{4.20}$$

Assumption 4.A $(\bar{e}_{\mathbf{A},k})_{k\in\mathbb{N}_0}$ is nonincreasing and converges to 0; $(n_k)_{k\in\mathbb{N}_0}$ is strictly increasing and $n_0 = 0$. Furthermore, the sequence $(\eta_k)_{k\in\mathbb{N}_0}$ is nonincreasing.

Note that Assumption 4.A implies Assumption 3.A. Let $\mathscr{M}$ denote the set of $p \in \{0, \dots, P\}$ for which $\operatorname{supp}\mathbf{v}_{[p]} \neq \varnothing$. For all $p \in \mathscr{M}$, the sequences of costs and values from Sect. 3 are given by

$$c_k^p := n_k(\#\operatorname{supp}\mathbf{v}_{[p]}) \qquad \text{and} \qquad \omega_k^p := -\bar{e}_{\mathbf{A},k}\|\mathbf{v}_{[p]}\|_{\ell^2}. \tag{4.21}$$

By Assumption 4.A, $c_0^p = 0$, $(c_k^p)_{k\in\mathbb{N}_0}$ is strictly increasing and $(\omega_k^p)_{k\in\mathbb{N}_0}$ is nondecreasing for all $p \in \mathscr{M}$. Also,

$$q_k^p = \frac{\Delta\omega_k^p}{\Delta c_k^p} = \eta_k \frac{\|\mathbf{v}_{[p]}\|_{\ell^2(\Xi_p)}}{\#\operatorname{supp}\mathbf{v}_{[p]}} \tag{4.22}$$

is nonincreasing in k for all $p \in \mathscr{M}$.

Proposition 4.6 *For any* $\mathbf{k}$ *generated in* $\mathtt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$, *if* $\mathbf{j} \in \mathbb{N}_0^{\ell}$ *with* $\sigma_{\mathbf{j}} \le \sigma_{\mathbf{k}}$, *then* $\zeta_{\mathbf{j}} \ge \zeta_{\mathbf{k}}$. *If* $\mathbf{j} \in \mathbb{N}_0^{\ell}$ *with* $\zeta_{\mathbf{j}} \le \zeta_{\mathbf{k}}$, *then* $\sigma_{\mathbf{j}} \ge \sigma_{\mathbf{k}}$.

Proof The assertion follows from Theorem 3.4 with (4.21) and using Assumption 4.A. Note that $\sigma_{\mathbf{j}} \ge 0$ for all $\mathbf{j} \in \mathbb{N}_0^{\ell}$, and if $\sigma_{\mathbf{k}} > 0$, the second statement in Theorem 3.4 applies.

Let $(\mathbf{k}_i)_{i\in\mathbb{N}_0}$ denote the sequence of $\mathbf{k}$ generated in $\mathtt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ if the loop is not terminated. We abbreviate $\zeta_i := \zeta_{\mathbf{k}_i}$ and $\sigma_i := \sigma_{\mathbf{k}_i}$.

Remark 4.7 In particular, Proposition 4.6 implies convergence of the greedy subroutine in $\mathtt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$. Since $n_{k+1} \ge n_k + 1$ for all $k \in \mathbb{N}_0$ and $k_{i,p} = 0$ for all $i \in \mathbb{N}_0$ if $\#\operatorname{supp}\mathbf{v}_{[p]} = 0$, σ_i goes to infinity as $i \to \infty$. Since $\zeta_{\mathbf{j}}$ can be made arbitrarily small for suitable $\mathbf{j} \in \mathbb{N}_0^{\ell}$, it follows that $\zeta_i \to 0$.

5 Analysis of the adaptive application routine

5.1 Convergence analysis

For the analysis of $\texttt{Apply}_{\mathbf{A}}$, we assume that the values $\bar{e}_{\mathbf{A},k}$ are spaced sufficiently regularly, with at most geometric convergence to 0.

Assumption 5.A $\bar{r}_{\mathbf{A}} := \sup_{k\in\mathbb{N}_0} \dfrac{\bar{e}_{\mathbf{A},k}}{\bar{e}_{\mathbf{A},k+1}} < \infty.$

In particular, $\bar{e}_{\mathbf{A},k} > 0$ for all $k \in \mathbb{N}_0$, i.e. if $\mathbf{A}$ is sparse, this is not reflected in the bounds $\bar{e}_{\mathbf{A},k}$. An admissible value is $\bar{e}_{\mathbf{A},k} = d_{\mathbf{A},s} n_k^{-s}$ since for all $k \in \mathbb{N}_0$,

$$\frac{\bar{e}_{\mathbf{A},k}}{\bar{e}_{\mathbf{A},k+1}} = \left(\frac{n_{k+1}}{n_k}\right)^s \leq c_{\mathbf{A}}^s < \infty.$$

Lemma 5.1 *For all $i \in \mathbb{N}_0$, $\zeta_i \leq \bar{r}_{\mathbf{A}} \zeta_{i+1}$.*

Proof Let $i \in \mathbb{N}_0$. Note that

$$\zeta_i - \zeta_{i+1} = (\bar{e}_{\mathbf{A},k_{q_i}} - \bar{e}_{\mathbf{A},k_{q_i}+1}) \|\mathbf{v}_{[q_i]}\|_{\ell^2} \quad \text{and} \quad \zeta_{i+1} \geq \bar{e}_{\mathbf{A},k_{q_i}+1} \|\mathbf{v}_{[q_i]}\|_{\ell^2}\,.$$

Therefore,

$$\frac{\zeta_i}{\zeta_{i+1}} = 1 + \frac{\zeta_i - \zeta_{i+1}}{\zeta_{i+1}} \leq 1 + \frac{\bar{e}_{\mathbf{A},k_{q_i}} - \bar{e}_{\mathbf{A},k_{q_i}+1}}{\bar{e}_{\mathbf{A},k_{q_i}+1}} = \frac{\bar{e}_{\mathbf{A},k_{q_i}}}{\bar{e}_{\mathbf{A},k_{q_i}+1}} \leq \bar{r}_{\mathbf{A}}.$$

The following theorem is adapted from [20, Thm. 4.6]. We emphasize in advance that knowledge of s and s^* is not required in $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$. The algorithm satisfies Theorem 5.2 with any s^* for which $\mathbf{A}$ is s^*-compressible, provided that the bounds $\bar{e}_{\mathbf{A},k}$ from (4.11) decay at the rate implied by s^*-compressibility. The constant in (5.2) may degenerate as $s \to s^*$.

Theorem 5.2 *Let $\mathbf{v} \in \ell^2$ be finitely supported and $\varepsilon > 0$. A call of $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ produces a finitely supported $\mathbf{z} \in \ell^2$ with*

$$\|\mathbf{A}\mathbf{v} - \mathbf{z}\|_{\ell^2} \leq \delta + \zeta_{\mathbf{k}} \leq \varepsilon. \tag{5.1}$$

If $\mathbf{A}$ is s^*-compressible for an $s^* \in (0, \infty]$ and $\sup_{k\in\mathbb{N}} \bar{e}_{\mathbf{A},k} n_k^s < \infty$ for all $s \in (0, s^*)$, then for any $s \in (0, s^*)$,

$$\#\operatorname{supp} \mathbf{z} \leq \sigma_{\mathbf{k}} \lesssim \varepsilon^{-1/s} \|\mathbf{v}\|_{\mathscr{A}^s}^{1/s} \tag{5.2}$$

with a constant depending only on s, $\bar{e}_{\mathbf{A},0}$, $c_{\mathbf{A}}$, n_1, $(d_{\mathbf{A},\bar{s}})_{\bar{s}\in(s,s^*)}$ and $\bar{r}_{\mathbf{A}}$.

Proof Convergence of $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ follows from Proposition 4.6, see Remark 4.7. Then (5.1) is shown in Proposition 4.5.

Let $\mathbf{k} = (k_p)_{p=1}^{\ell}$ be the final value of $\mathbf{k}$ in $\mathtt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$, and $s \in (0, s^*)$. By Proposition 4.5, to prove (5.2) it suffices to show that there is a $\mathbf{j} \in \mathbb{N}_0^{\ell}$ with $\zeta_{\mathbf{j}} \le \zeta_{\mathbf{k}} =: \zeta$ and $\sigma_{\mathbf{j}} \lesssim \varepsilon^{-1/s} \|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}$. Then Proposition 4.6 implies

$$\#\operatorname{supp} \mathbf{z} \le \sigma_{\mathbf{k}} \le \sigma_{\mathbf{j}} \lesssim \varepsilon^{-1/s} \|\mathbf{v}\|_{\mathscr{A}^s}^{1/s} .$$

The construction of such a $\mathbf{j}$ is analogous to the proof of [20, Thm. 4.6] with ζ in place of $\varepsilon - \delta$. We provide it here for completeness.

Let $\tau \in (0, 2)$ be defined by $\tau^{-1} = s + \frac{1}{2}$, and let $s < \bar{s}_1 < \bar{s}_2 < s^*$. Then

$$\#\operatorname{supp} \mathbf{v}_{[p]} \le \# \left\{ \mu \in \Xi \,;\, \left|v_\mu\right| > 2^{-p/2} \|\mathbf{v}\|_{\ell^\infty} \right\} \lesssim 2^{p\tau/2} \|\mathbf{v}\|_{\ell^\infty}^{-\tau} \|\mathbf{v}\|_{\mathscr{A}^s}^{\tau},$$

see e.g. [19]. In particular,

$$\|\mathbf{v}_{[p]}\|_{\ell^2} \le 2^{-(p-1)/2} \|\mathbf{v}\|_{\ell^\infty} \sqrt{\#\operatorname{supp} \mathbf{v}_{[p]}} \lesssim 2^{-ps\tau/2} \|\mathbf{v}\|_{\ell^\infty}^{1-\tau/2} \|\mathbf{v}\|_{\mathscr{A}^s}^{\tau/2}.$$

Let $J \ge \ell$ be the smallest integer with $\sum_{p=1}^{\ell} 2^{-(J-p)\bar{s}_1\tau/2} \|\mathbf{v}_{[p]}\|_{\ell^2} \le \zeta$ and let $\mathbf{j} = (j_p)_{p=1}^{\ell} \in \mathbb{N}_0^{\ell}$ with j_p minimal such that $\bar{e}_{\mathbf{A}, j_p} \le 2^{-(J-p)\bar{s}_1\tau/2}$. Then

$$\zeta_{\mathbf{j}} = \sum_{p=1}^{\ell} \bar{e}_{\mathbf{A}, j_p} \|\mathbf{v}_{[p]}\|_{\ell^2} \le \sum_{p=1}^{\ell} 2^{-(J-p)\bar{s}_1\tau/2} \|\mathbf{v}_{[p]}\|_{\ell^2} \le \zeta.$$

It remains to be shown that $\sigma_{\mathbf{j}} \lesssim \varepsilon^{-1/s} \|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}$.

If $j_p \ge 2$, since $\bar{e}_{\mathbf{A}, j_p - 1} n_{j_p-1}^{\bar{s}_2} \lesssim 1$,

$$n_{j_p} \lesssim n_{j_p-1} \lesssim \bar{e}_{\mathbf{A}, j_p-1}^{-1/\bar{s}_2} \le 2^{(J-p)(\bar{s}_1/\bar{s}_2)\tau/2}.$$

This estimate extends to $j_p \in \{0, 1\}$ since $p \le J$. Therefore, using $\bar{s}_1 < \bar{s}_2$,

$$\begin{aligned} \sigma_{\mathbf{j}} &= \sum_{p=1}^{\ell} n_{j_p} (\#\operatorname{supp} \mathbf{v}_{[p]}) \lesssim \sum_{p=1}^{\ell} 2^{(J-p)(\bar{s}_1/\bar{s}_2)\tau/2} 2^{-ps\tau/2} \|\mathbf{v}\|_{\ell^\infty}^{-\tau} \|\mathbf{v}\|_{\mathscr{A}^s}^{\tau} \\ &\lesssim 2^{(J-\ell)(\bar{s}_1/\bar{s}_2)\tau/2} 2^{-\ell s\tau/2} \|\mathbf{v}\|_{\ell^\infty}^{-\tau} \|\mathbf{v}\|_{\mathscr{A}^s}^{\tau} \le 2^{J\tau/2} \|\mathbf{v}\|_{\ell^\infty}^{-\tau} \|\mathbf{v}\|_{\mathscr{A}^s}^{\tau}. \end{aligned}$$

Thus, the assertion reduces to $2^{J\tau/2} \|\mathbf{v}\|_{\ell^\infty}^{-\tau} \|\mathbf{v}\|_{\mathscr{A}^s}^{\tau} \lesssim \varepsilon^{-1/s} \|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}$.

If $J = \ell$, by minimality of ℓ,

$$\frac{\varepsilon}{2} < \bar{e}_{\mathbf{A},0} \left\| \mathbf{v} - \sum_{p=1}^{\ell-1} \mathbf{v}_{[p]} \right\| = \bar{e}_{\mathbf{A},0} \sqrt{\sum_{p=\ell}^{\infty} \|\mathbf{v}_{[p]}\|_{\ell^2}^2} \lesssim \bar{e}_{\mathbf{A},0} 2^{-\ell s\tau/2} \|\mathbf{v}\|_{\ell^\infty}^{1-\tau/2} \|\mathbf{v}\|_{\mathscr{A}^s}^{\tau/2} .$$

If $J > \ell$, then by minimality of J, using $s < \bar{s}_1$,

$$\zeta < \sum_{p=1}^{\ell} 2^{-(J-1-p)\bar{s}_1\tau/2}\|\mathbf{v}_{[p]}\|_{\ell^2} \lesssim \sum_{p=1}^{\ell} 2^{-(J-1-p)\bar{s}_1\tau/2}2^{-ps\tau/2}\|\mathbf{v}\|_{\ell^\infty}^{1-\tau/2}\|\mathbf{v}\|_{\mathscr{A}^s}^{\tau/2}$$
$$\lesssim 2^{-(J-1-\ell)\bar{s}_1\tau/2}2^{-\ell s\tau/2}\|\mathbf{v}\|_{\ell^\infty}^{1-\tau/2}\|\mathbf{v}\|_{\mathscr{A}^s}^{\tau/2} \leq 2^{-(J-1)s\tau/2}\|\mathbf{v}\|_{\ell^\infty}^{1-\tau/2}\|\mathbf{v}\|_{\mathscr{A}^s}^{\tau/2}.$$

Lemma 5.1 implies $\varepsilon \leq \bar{r}_{\mathbf{A}}\zeta$. Therefore, in both cases,

$$\varepsilon \lesssim 2^{-Js\tau/2}\|\mathbf{v}\|_{\ell^\infty}^{1-\tau/2}\|\mathbf{v}\|_{\mathscr{A}^s}^{\tau/2},$$

or equivalently,

$$2^{J\tau/2}\|\mathbf{v}\|_{\ell^\infty}^{-\tau}\|\mathbf{v}\|_{\mathscr{A}^s}^{\tau} \lesssim \varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s},$$

which completes the proof.

It is known that s^*-compressible operators $\mathbf{A}$ map $\mathscr{A}^s$ boundedly into $\mathscr{A}^s$ for all $s \in (0, s^*)$, see [9, Proposition 3.8]. Theorem 5.2 implies that this carries over to the approximate multiplication routine $\texttt{Apply}_{\mathbf{A}}$.

Corollary 5.3 *Let* $\mathbf{A}$ *be* s^**-compressible for some* $s^* \in (0, \infty]$*, and assume that for all* $s \in (0, s^*)$*,* $\sup_{k\in\mathbb{N}} \bar{e}_{\mathbf{A},k}n_k^s < \infty$*. Then for any* $s \in (0, s^*)$ *there is a constant* C *depending only on* s*,* $\bar{e}_{\mathbf{A},0}$*,* $c_{\mathbf{A}}$*,* n_1*,* $(d_{\mathbf{A},\bar{s}})_{\bar{s}\in(s,s^*)}$ *and* $\bar{r}_{\mathbf{A}}$ *such that for all* $\mathbf{v} \in \mathscr{A}^s$ *and all* $\varepsilon > 0$*, the output* $\mathbf{z}$ *of* $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ *satisfies*

$$\|\mathbf{z}\|_{\mathscr{A}^s} \leq C\|\mathbf{v}\|_{\mathscr{A}^s}. \tag{5.3}$$

Proof Let $\mathbf{z}$ be the output of $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ for some $\mathbf{v} \in \mathscr{A}^s$ and some $\varepsilon > 0$, and define $\mathbf{w} := \mathbf{A}\mathbf{v}$. By [9, Proposition 3.8], $\mathbf{w} \in \mathscr{A}^s$, and $\|\mathbf{w}\|_{\mathscr{A}^s} \lesssim \|\mathbf{v}\|_{\mathscr{A}^s}$. Since $\mathbf{z}$ is finitely supported, $\mathbf{z} \in \mathscr{A}^s$. Let $N := \#\operatorname{supp}\mathbf{z}$. Theorem 5.2 implies

$$\|\mathbf{w} - \mathbf{z}\|_{\ell^2} \lesssim \|\mathbf{v}\|_{\mathscr{A}^s}N^{-s}.$$

For any $n \geq N$, $P_n(\mathbf{z}) = \mathbf{z}$, and thus $(n+1)^s\|\mathbf{z} - P_n(\mathbf{z})\|_{\ell^2} = 0$. Let $n \leq N-1$ and $\mathbf{z}_n \in \ell^2$ with $\#\operatorname{supp}\mathbf{z}_n \leq n$. Then

$$(n+1)^s\|\mathbf{z} - \mathbf{z}_n\|_{\ell^2} \leq (n+1)^s\|\mathbf{w} - \mathbf{z}\|_{\ell^2} + (n+1)^s\|\mathbf{w} - \mathbf{z}_n\|_{\ell^2}.$$

The first term is bounded by

$$(n+1)^s\|\mathbf{w} - \mathbf{z}\|_{\ell^2} \lesssim (n+1)^sN^{-s}\|\mathbf{v}\|_{\mathscr{A}^s} \lesssim \|\mathbf{v}\|_{\mathscr{A}^s}.$$

Taking the infimum over $\mathbf{z}_n$ with $\#\operatorname{supp}\mathbf{z}_n \leq n$ and using $\|\mathbf{w}\|_{\mathscr{A}^s} \lesssim \|\mathbf{v}\|_{\mathscr{A}^s}$, we have

$$(n+1)^s\|\mathbf{z} - P_n(\mathbf{z})\|_{\ell^2} \lesssim \|\mathbf{v}\|_{\mathscr{A}^s} + (n+1)^s\inf_{\mathbf{z}_n}\|\mathbf{w} - \mathbf{z}_n\|_{\ell^2} \lesssim \|\mathbf{v}\|_{\mathscr{A}^s}.$$

The assertion follows by taking the supremum over $n \in \mathbb{N}_0$.

5.2 Complexity analysis

By (4.15), the number of operations and storage locations required by `BucketSort` in a call of $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ is bounded by

$$\begin{aligned} &\#\operatorname{supp}\mathbf{v} + \max(1, \lceil \log(2\bar{e}_{\mathbf{A},0}\|\mathbf{v}\|_{\ell^\infty}\sqrt{\#\operatorname{supp}\mathbf{v}}/\varepsilon)\rceil) \\ &\quad\lesssim 1 + \#\operatorname{supp}\mathbf{v} + \log(\varepsilon^{-1}\|\mathbf{v}\|_{\ell^\infty}). \end{aligned} \tag{5.4}$$

The value of ℓ can be determined with at most $\#\operatorname{supp}\mathbf{v}$ operations. We assume that the values of $\|\mathbf{v}_{[p]}\|_{\ell^2(\Xi_p)}$ are known from the computation of ℓ. Then by Proposition 3.5, initialization of the greedy subroutine requires $\mathcal{O}(\ell \log \ell)$ operations, and each iteration requires $\mathcal{O}(1 + \log \ell)$ operations e.g. if a tree data structure is used for $\mathcal{N}$ from Sect. 3.3. As $\|\mathbf{k}\|_{\ell^1}$ iterations are performed if $\mathbf{k} = (k_p)_{p=1}^{\ell}$ is the final value of $\mathbf{k}$ in $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$, the total cost of determining ℓ and $\mathbf{k}$ is on the order of

$$\#\operatorname{supp}\mathbf{v} + \ell \log^+ \ell + (1 + \log^+ \ell) \sum_{p=1}^{\ell} k_p, \tag{5.5}$$

where $\log^+ x := \log(\max(x, 1))$. Since $\ell \le P$, (4.14) implies

$$\ell \lesssim 1 + \log^+(\#\operatorname{supp}\mathbf{v}) + \log^+(\varepsilon^{-1}\|\mathbf{v}\|_{\ell^\infty}). \tag{5.6}$$

Finally, the number of arithmetic operations required by the last step of $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ is bounded by

$$\sigma_{\mathbf{k}} = \sum_{p=1}^{\ell} n_{k_p}(\#\operatorname{supp}\mathbf{v}_{[p]}), \tag{5.7}$$

and this value is optimal in the sense of Proposition 4.6. If $\mathbf{A}$ is s^*-computable for any $s^* \in (0, \infty]$, then (5.7) includes the assembly costs of $\mathbf{A}_{k_p}$.

Theorem 5.4 *Let $\mathbf{v} \in \ell^2$ be finitely supported and $\varepsilon > 0$. If $\mathbf{A}$ is s^*-computable for an $s^* \in (0, \infty]$ and $\sup_{k\in\mathbb{N}} \bar{e}_{\mathbf{A},k} n_k^s < \infty$ for all $s \in (0, s^*)$, then for any $s \in (0, s^*)$, the number of operations and storage locations required by $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ is less than a multiple of*

$$1 + \#\operatorname{supp}\mathbf{v} + \varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}\left(1 + \log^+\log^+\left(\#\operatorname{supp}\mathbf{v} + \varepsilon^{-1}\|\mathbf{v}\|_{\ell^\infty}\right)\right) \tag{5.8}$$

with a constant depending only on s, $\bar{e}_{\mathbf{A},0}$, $c_{\mathbf{A}}$, n_1, $(d_{\mathbf{A},\bar{s}})_{\bar{s}\in(s,s^*)}$, $\bar{r}_{\mathbf{A}}$ and $b_{\mathbf{A}}$. The double logarithmic term in (5.8) is due only to the greedy subroutine and does not apply to the storage requirements.[1]

[1] As above, $\log^+ x := \log(\max(x, 1))$.

Proof We first note that

$$\log(\varepsilon^{-1}\|\mathbf{v}\|_{\ell^\infty}) \lesssim \varepsilon^{-1/s}\|\mathbf{v}\|_{\ell^\infty}^{1/s} \leq \varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}.$$

Therefore and by (5.4), the cost of `BucketSort` is less than

$$1 + \#\operatorname{supp}\mathbf{v} + \log(\varepsilon^{-1}\|\mathbf{v}\|_{\ell^\infty}) \lesssim 1 + \#\operatorname{supp}\mathbf{v} + \varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}.$$

The cost of the last step of $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ is $\sigma_{\mathbf{k}}$, which in Theorem 5.2 is bounded by

$$\sigma_{\mathbf{k}} \lesssim \varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}.$$

The cost of the rest of $\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \varepsilon]$ is given in (5.5). By (5.6), for $\chi > 1$,

$$\begin{aligned}\ell \log \ell &\lesssim \ell^\chi \lesssim 1 + \log(\#\operatorname{supp}\mathbf{v})^\chi + \log(\varepsilon^{-1}\|\mathbf{v}\|_{\ell^\infty})^\chi \\ &\lesssim 1 + \#\operatorname{supp}\mathbf{v} + \varepsilon^{-1/s}\|\mathbf{v}\|_{\ell^\infty}^{1/s} \leq 1 + \#\operatorname{supp}\mathbf{v} + \varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}.\end{aligned}$$

Since

$$\ell \lesssim 1 + \log(\#\operatorname{supp}\mathbf{v}) + \log(\varepsilon^{-1}\|\mathbf{v}\|_{\ell^\infty}) \lesssim 1 + \log(\#\operatorname{supp}\mathbf{v} + \varepsilon^{-1}\|\mathbf{v}\|_{\ell^\infty}),$$

we have

$$\log \ell \leq C + \log(1 + \log(\#\operatorname{supp}\mathbf{v} + \varepsilon^{-1}\|\mathbf{v}\|_{\ell^\infty})) \lesssim 1 + \log\log(\#\operatorname{supp}\mathbf{v} + \varepsilon^{-1}\|\mathbf{v}\|_{\ell^\infty}).$$

Finally, since $k \leq n_k$ for all $k \in \mathbb{N}_0$ and $k_p = 0$ if $\#\operatorname{supp}\mathbf{v}_{[p]} = 0$,

$$\sum_{p=1}^{\ell} k_p \leq \sum_{p=1}^{\ell} n_{k_p}(\#\operatorname{supp}\mathbf{v}_{[p]}) = \sigma_{\mathbf{k}} \lesssim \varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}.$$

Remark 5.5 The double logarithmic term in (5.8) can be dropped under mild conditions. If $n_k \gtrsim k^\alpha$ for an $\alpha > 1$, then by Hölder's inequality,

$$\sum_{p=1}^{\ell} k_p \lesssim \sum_{p=1}^{\ell} n_{k_p}^{1/\alpha} \leq \Big(\sum_{p=1}^{\ell} n_{k_p}\Big)^{1/\alpha} \ell^{\frac{\alpha-1}{\alpha}}.$$

Furthermore, for a $\chi > 1$, as in the proof of Theorem 5.4,

$$\ell^{\frac{\alpha-1}{\alpha}} \log \ell \lesssim (\ell^\chi)^{\frac{\alpha-1}{\alpha}} \lesssim \big(1 + \#\operatorname{supp}\mathbf{v} + \varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}\big)^{\frac{\alpha-1}{\alpha}}.$$

It follows that

$$\log \ell \sum_{p=1}^{\ell} k_p \lesssim \sigma_{\mathbf{k}}^{1/\alpha}\Big(1 + \#\operatorname{supp}\mathbf{v} + \varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s}\Big)^{\frac{\alpha-1}{\alpha}} \lesssim 1 + \#\operatorname{supp}\mathbf{v} + \varepsilon^{-1/s}\|\mathbf{v}\|_{\mathscr{A}^s}^{1/s},$$

and (5.8) can be replaced by

$$1 + \#\operatorname{supp} \mathbf{v} + \varepsilon^{-1/s} \|\mathbf{v}\|_{\mathscr{A}^s}^{1/s} \tag{5.9}$$

in Theorem 5.4, with a constant that also depends on α. The assumption $n_k \gtrsim k^\alpha$ is generally not restrictive, since by (4.1), n_k may grow exponentially for an s^*-compressible operator.

6 Computation of spectral norms by the power method

6.1 Estimation of errors in sparse approximations of s^*-compressible operators

The routine $\texttt{Apply}_{\mathbf{A}}$ in Sect. 4.2 makes explicit use of bounds $\bar{e}_{\mathbf{A},k}$ on the errors $\|\mathbf{A} - \mathbf{A}_k\|_{\ell^2 \to \ell^2}$, where $\mathbf{A}_k$ is an n_k-sparse approximation of an operator $\mathbf{A} \in \mathscr{L}(\ell^2)$, see (4.11). Such bounds are derived e.g. in [3,34] for a large class of operators in wavelet bases. However, these estimates only hold up to an unspecified constant.

We suggest a power method for numerically approximating $\|\mathbf{A} - \mathbf{A}_k\|_{\ell^2 \to \ell^2}$, which is equal to the square root of the spectral radius of the bounded positive symmetric operator $(\mathbf{A} - \mathbf{A}_k)^*(\mathbf{A} - \mathbf{A}_k)$ on ℓ^2.

Remark 6.1 If $\mathbf{A}$ is s^*-compressible with a sequence $(\mathbf{A}_j)_{j\in\mathbb{N}}$ of n_j-sparse approximations, then $\mathbf{A} - \mathbf{A}_k$ is also s^*-compressible with approximations $(\mathbf{A}_{k+j} - \mathbf{A}_k)_{j\in\mathbb{N}}$. We have

$$\|(\mathbf{A} - \mathbf{A}_k) - (\mathbf{A}_{k+j} - \mathbf{A}_k)\|_{\ell^2 \to \ell^2} = \|\mathbf{A} - \mathbf{A}_{k+j}\|_{\ell^2 \to \ell^2} = e_{\mathbf{A},k+j} \le d_{\mathbf{A},s} n_{k+j}^{-s}. \tag{6.1}$$

Furthermore, $\mathbf{A}_{k+j} - \mathbf{A}_k$ is at most $(n_{k+j} + n_k)$-sparse, which implies $d_{\mathbf{A}-\mathbf{A}_k,s} \le 2^s d_{\mathbf{A},s}$. If the nonzero entries of $\mathbf{A}_k$ are also nonzero for $\mathbf{A}_{k+j}$, then $\mathbf{A}_{k+j} - \mathbf{A}_k$ is n_{k+j}-sparse, or even $(n_{k+j} - n_k)$-sparse if the values of these entries coincide. In either case, $d_{\mathbf{A}-\mathbf{A}_k,s} \le d_{\mathbf{A},s}$. Similar considerations lead to $c_{\mathbf{A}-\mathbf{A}_k} \le 2c_{\mathbf{A}}^2/(c_{\mathbf{A}} - 1)$.

6.2 Analysis of an idealized iteration

Let $\mathbf{A} \in \mathscr{L}(\ell^2)$ be a positive symmetric operator. The power method successively approximates the spectral radius $r_{\mathbf{A}}$ of $\mathbf{A}$ by Rayleigh quotients

$$R_n := \frac{\left(\mathbf{A}^{n+1}\mathbf{v}, \mathbf{A}^n\mathbf{v}\right)_{\ell^2}}{\|\mathbf{A}^n\mathbf{v}\|_{\ell^2}^2} = \frac{\left(\mathbf{A}^{2n+1}\mathbf{v}, \mathbf{v}\right)_{\ell^2}}{\left(\mathbf{A}^{2n}\mathbf{v}, \mathbf{v}\right)_{\ell^2}}, \quad n \in \mathbb{N}, \tag{6.2}$$

for some starting value $\mathbf{v} \in \ell^2$.

Remark 6.2 The classical analysis of the power method in a finite dimensional setting makes use of the gap between the two largest eigenvalues. In our infinite dimensional setting, such a gap need not exist, and thus a different analysis is required.

Theorem 6.3 *For appropriate starting values* $\mathbf{v} \in \ell^2$ *and any* $\vartheta \in (0,1)$*, there is a constant* $c_{\mathbf{v},\vartheta} > 0$ *such that*

$$r_{\mathbf{A}} \geq R_n \geq \vartheta r_{\mathbf{A}}(1 - c_{\mathbf{v},\vartheta} n^{-1}) \qquad \forall n \in \mathbb{N}\,. \tag{6.3}$$

In particular, $R_n \to r_{\mathbf{A}}$.

Proof We note that $R_n \leq r_{\mathbf{A}}$ for all $n \in \mathbb{N}$ by definition. Due to the spectral theorem for bounded symmetric operators, there is a σ-finite measure μ on some domain S and a unitary map $U\colon L^2_\mu(S) \to \ell^2$ such that

$$U^*\mathbf{A}U\varphi = f\varphi \qquad \forall \varphi \in L^2_\mu(S)\,,$$

where $f \in L^\infty_\mu(S)$ with $f \geq 0$ and $r_{\mathbf{A}} = \|f\|_{L^\infty_\mu(S)}$. We assume without loss of generality that $\|\mathbf{v}\|_{\ell^2} = 1$ and define $\varphi := U^*\mathbf{v}$. Then the Rayleigh quotients (6.2) are

$$R_n = \frac{\int_S f^{2n+1}|\varphi|^2\,\mathrm{d}\mu}{\int_S f^{2n}|\varphi|^2\,\mathrm{d}\mu} = \frac{\int_S f^{2n+1}\,\mathrm{d}\mu_\varphi}{\int_S f^{2n}\,\mathrm{d}\mu_\varphi}$$

for the probability measure $\mathrm{d}\mu_\varphi := |\varphi|^2\,\mathrm{d}\mu$. By Jensen's inequality, $\|f\|_{L^{2n+1}_{\mu_\varphi}(S)} \geq \|f\|_{L^{2n}_{\mu_\varphi}(S)}$, and thus

$$R_n \geq \frac{\left(\int_S f^{2n}\,\mathrm{d}\mu_\varphi\right)^{\frac{2n+1}{2n}}}{\int_S f^{2n}\,\mathrm{d}\mu_\varphi} = \left(\int_S f^{2n}\,\mathrm{d}\mu_\varphi\right)^{\frac{1}{2n}} = \|f\|_{L^{2n}_{\mu_\varphi}(S)}.$$

Since $\|f\|_{L^p_{\mu_\varphi}(S)} \to \|f\|_{L^\infty_{\mu_\varphi}(S)}$ as $p \to \infty$, convergence of R_n to $r_{\mathbf{A}}$ follows, provided that

$$\operatorname*{ess\,sup}_{x \in \operatorname{supp}\varphi} f(x) = \operatorname*{ess\,sup}_{x \in S} f(x). \tag{6.4}$$

We estimate $\|f\|_{L^{2n}_{\mu_\varphi}(S)}$ from below in order to get a convergence rate. Let $\vartheta \in (0,1)$. Then Markov's inequality implies

$$\|f\|_{L^{2n}_{\mu_\varphi}(S)} \geq \vartheta \|f\|_{L^\infty_{\mu_\varphi}(S)} \kappa^{1/2n}, \quad \kappa := \mu_\varphi\left(\left\{x \in S\,;\ f(x) \geq \vartheta\|f\|_{L^\infty_{\mu_\varphi}(S)}\right\}\right) \in (0,1].$$

Furthermore, by the fundamental theorem of calculus,

$$\kappa^{1/2n} \geq 1 - (1-\kappa)\frac{1}{2n}\kappa^{\frac{1}{2n}-1} \geq 1 - \frac{1-\kappa}{2\kappa}\frac{1}{n}.$$

The proof of Theorem 6.3 clarifies the conditions on the starting value $\mathbf{v}$: It must satisfy (6.4) for $\varphi = U^*\mathbf{v}$ and f as in the proof, which is analogous to the assumption that the starting vector in a finite dimensional power method is not orthogonal to the

eigenspace associated to the largest eigenvalue. We expect round-off errors to make this condition irrelevant for numerical computations.

6.3 A practical algorithm

The Rayleigh quotients (6.2) cannot be computed exactly since the operator $\mathbf{A}$ cannot by applied exactly. We suggest an approximate adaptive procedure for evaluating $\mathbf{Av}$ similar to the routine $\texttt{Apply}_{\mathbf{A}}$ from Sect. 4.2. To this end, we assume that for all $k \in \mathbb{N}_0$, $\mathbf{A}_k$ is an n_k-sparse approximation of $\mathbf{A}$, with $n_0 = 0$, $n_{k+1} \geq n_k + 1$ for all $k \in \mathbb{N}_0$ and

$$\|\mathbf{A} - \mathbf{A}_k\|_{\ell^2 \to \ell^2} \leq C \tilde{e}_{\mathbf{A},k} \tag{6.5}$$

for a constant C. We emphasize that this assumption is weaker than (4.11) since the constant C need not be known, and our algorithm does not depend on this constant. If it is known that $\mathbf{A}$ is s^*-compressible, then we may set $\tilde{e}_{\mathbf{A},k} := n_k^{-s}$ for any $s \in (0, s^*)$.

Let $\mathbf{v} = (v_\mu)_{\mu \in \mathbb{N}}$ be a finitely supported sequence. We consider a sorting routine

$$\texttt{Sort}[\mathbf{v}] \mapsto (\mu_i)_{i=1}^M \tag{6.6}$$

with $M := \#\operatorname{supp} \mathbf{v}$ and such that $(|v_{\mu_i}|)_{i=1}^M$ is a decreasing rearrangement of $(|v_\mu|)_{\mu \in \mathbb{N}}$. To approximate $\mathbf{Av}$, we apply either $\mathbf{A}_k$ or a better approximation of $\mathbf{A}$ to the first m_k terms of this decreasing rearrangement, i.e. we apply $\mathbf{A}_k$ to $\mathbf{v}$ restricted to the set $\{\mu_i \,;\, m_{k+1} + 1 \leq i \leq m_k\}$. For any nonincreasing sequence $\mathbf{m} = (m_k)_{k=1}^\infty$, the number of multiplications performed in this approximate application of $\mathbf{A}$ is at most

$$\sigma_{\mathbf{m}} := \sum_{k=1}^\infty n_k (m_k - m_{k+1}) = \sum_{k=1}^\infty (n_k - n_{k-1}) m_k, \tag{6.7}$$

and the error is bounded by

$$\chi_{\mathbf{m}} := \sum_{k=1}^\infty \tilde{e}_{\mathbf{A},k} \left(\sum_{i=m_{k+1}+1}^{m_k} |v_{\mu_i}|^2 \right)^{1/2}. \tag{6.8}$$

The routine $\texttt{NextOptInf}$ from Sect. 3.3 extends to objectives of the form $\chi_{\mathbf{m}}$ in a straightforward manner, and its output $\mathbf{m}$ is assured to be nonincreasing.

The routine $\texttt{NApply}_{\mathbf{A}}$ does not ensure a fixed error, contrary to $\texttt{Apply}_{\mathbf{A}}$. This would not be possible due to the unknown constant in the estimate (6.5). Instead, $\texttt{NApply}_{\mathbf{A}}$ limits the computational cost of the approximate multiplication. It can be thought of as an adaptively constructed matrix representation of $\mathbf{A}$ of size $N \times M$.

Remark 6.4 By construction, $\sigma_{\mathbf{m}} \leq N$ for the final value of $\mathbf{m}$ in $\texttt{NApply}_{\mathbf{A}}$. This implies that no more than N multiplications are performed in the computation of $\mathbf{z}$ in the final step of $\texttt{NApply}_{\mathbf{A}}$, and thus $\#\operatorname{supp} \mathbf{z} \leq N$.

$\texttt{NApply}_{\mathbf{A}}[\mathbf{v}, N] \mapsto \mathbf{z}$

$(\mu_i)_{i=1}^{M} \longleftarrow \texttt{Sort}[\mathbf{v}]$
$\mathbf{m} = (m_k)_{k=1}^{\infty} \longleftarrow (0)_{k=1}^{\infty}$
$\hat{\mathbf{m}} = (\hat{m}_k)_{k=1}^{\infty} \longleftarrow (0)_{k=1}^{\infty}$
while $\sigma_{\hat{\mathbf{m}}} \leq N$ **do**
  $\mathbf{m} \longleftarrow \hat{\mathbf{m}}$
  $\hat{\mathbf{m}} \longleftarrow \texttt{NextOptInf}[\mathbf{m}]$ with objective $-\chi_{\mathbf{m}}$ and cost $\sigma_{\mathbf{m}}$
forall the $k \in \mathbb{N}$ **do** $\Xi_k \longleftarrow \{\mu_i \,;\, m_{k+1} + 1 \leq i \leq m_k\}$
$\mathbf{z} \longleftarrow \sum_{k=1}^{\infty} \mathbf{A}_k \mathbf{v}|_{\Xi_k}$

Remark 6.5 The exact sorting in the first step of $\texttt{NApply}_{\mathbf{A}}$ uses $\mathcal{O}(M \log M)$ operations. If n_k increases exponentially in k and $\tilde{e}_{\mathbf{A},k}$ decreases exponentially in k, then at most $\mathcal{O}(N \log N)$ steps are required in the subsequent greedy algorithm. By Proposition 3.6, these can be realized at a computational cost of $\mathcal{O}(N(\log N)^2)$. Finally, as noted in Remark 6.4, the actual computation of $\mathbf{z}$ uses $\mathcal{O}(N)$ operations.

Starting from an arbitrary finitely supported nonzero $\mathbf{v} \in \ell^2$, $\texttt{SpecRad}_{\mathbf{A}}$ iteratively uses $\texttt{NApply}_{\mathbf{A}}$ to approximate multiplications by $\mathbf{A}$ in the Rayleigh quotients (6.2). As a termination criterion, lacking alternatives, we simply compare two consecutive approximations of the spectral radius of $\mathbf{A}$.

$\texttt{SpecRad}_{\mathbf{A}}[\mathbf{v}, N, \varepsilon] \mapsto \rho$

$\rho \longleftarrow \infty$
$\mathbf{v} \longleftarrow \mathbf{v}/\|\mathbf{v}\|_{\ell^2}$
repeat
  $\rho_0 \longleftarrow \rho$
  $\mathbf{w} \longleftarrow \texttt{NApply}_{\mathbf{A}}[\mathbf{v}, N]$
  $\rho \longleftarrow \mathbf{w} \cdot \mathbf{v}$
  $\mathbf{v} \longleftarrow \mathbf{w}/\|\mathbf{w}\|_{\ell^2}$
until $|\rho - \rho_0| \leq \varepsilon\rho$

Remark 6.6 Since N is held constant throughout $\texttt{SpecRad}_{\mathbf{A}}$, assuming $\#\operatorname{supp}\mathbf{v} \leq N$ for the starting value of $\mathbf{v}$, each step of $\texttt{SpecRad}_{\mathbf{A}}$ has a computational cost of $\mathcal{O}(N(\log N)^2)$ due to Remark 6.5. Consequently, the choice of $\mathbf{v}$ is not particularly important—a poor choice is likely to be compensated by a few steps of the iteration, and the cost of subsequent steps is not affected. Note that the situation would be different if $\texttt{Apply}_{\mathbf{A}}$ were used in place of $\texttt{NApply}_{\mathbf{A}}$.

Remark 6.7 In order to compute the spectral radius of $\mathbf{A}^*\mathbf{A}$ for an operator $\mathbf{A} \in \mathscr{L}(\ell^2)$ that is not positive, instead of constructing sparse approximations of $\mathbf{A}^*\mathbf{A}$, the algorithm $\texttt{SpecRad}_{\mathbf{A}^*\mathbf{A}}$ can be used with $\texttt{NApply}_{\mathbf{A}^*\mathbf{A}}[\mathbf{v}, N]$ replaced by

$$\texttt{NApply}_{\mathbf{A}^*}[\texttt{NApply}_{\mathbf{A}}[\mathbf{v}, N], N]. \tag{6.9}$$

All vectors appearing in the iteration are still ensured to have at most N nonzero entries, and Remark 6.6 still holds. This can be used in the setting of Sect. 6.1, with $\mathbf{A} - \mathbf{A}_k$ in place of $\mathbf{A}$.

7 Random operator equations

7.1 Pathwise definition

Let V be a real separable Hilbert space, and V^* its dual space. We consider operator equations depending on a parameter in $\Gamma := [-1, 1]^\infty$. Given

$$A\colon \Gamma \to \mathscr{L}(V, V^*) \qquad \text{and} \qquad f\colon \Gamma \to V^*, \tag{7.1}$$

such that the bilinear form $\langle A(y)\cdot, \cdot\rangle$ is symmetric and positive on V for all $y \in \Gamma$, we wish to determine

$$u\colon \Gamma \to V, \qquad A(y)u(y) = f(y) \quad \forall y \in \Gamma. \tag{7.2}$$

Let $\mathscr{B}(\Gamma)$ denote the Borel σ-algebra on Γ. Defining a probability measure π on $(\Gamma, \mathscr{B}(\Gamma))$, A, f and u become random variables. Although π is arbitrary in this section, we assume in Sect. 7.2 below that π is a countable product of probability measures on $[-1, 1]$.

We decompose the operator A into deterministic and random components,

$$A(y) = D + R(y) \qquad \forall y \in \Gamma, \tag{7.3}$$

with $D \in \mathscr{L}(V, V^*)$ boundedly invertible and $R(y) \in \mathscr{L}(V, V^*)$ for all $y \in \Gamma$. Consequently, we also have the multiplicative decomposition

$$A(y) = D\big(\mathrm{id}_V + D^{-1}R(y)\big), \quad y \in \Gamma. \tag{7.4}$$

Under the assumption

$$\|D^{-1}R(y)\|_{V\to V} \le \gamma < 1 \qquad \forall y \in \Gamma, \tag{7.5}$$

a Neumann series argument ensures existence and uniqueness of the solution $u(y)$ of (7.2) for all $y \in \Gamma$, and

$$\|A(y)\|_{V\to V^*} \le \|D\|_{V\to V^*}(1 + \gamma) \qquad \forall y \in \Gamma, \tag{7.6}$$

$$\|A(y)^{-1}\|_{V^*\to V} \le \frac{1}{1-\gamma}\|D^{-1}\|_{V^*\to V} \qquad \forall y \in \Gamma. \tag{7.7}$$

As in e.g. [5,6,12,36], we consider random components that are linear in $y \in \Gamma$,

$$R(y) = \sum_{m=1}^{\infty} y_m R_m \qquad \forall y = (y_m)_{m=1}^\infty \in \Gamma, \tag{7.8}$$

with $R_m \in \mathscr{L}(V, V^*)$ for all m. Such operators arise e.g. if A is a differential operator that depends affinely on a random field and this field is expanded in a series. We assume that $(R_m)_m \in \ell^1(\mathbb{N}; \mathscr{L}(V, V^*))$ with

$$\sum_{m=1}^{\infty} \|D^{-1} R_m\|_{V\to V} \leq \gamma < 1, \tag{7.9}$$

which implies (7.5) since $|y_m| \leq 1$.

7.2 Discretization

Let the map $\Gamma \ni y \mapsto A(y)w(y)$ be measurable for any measurable $w\colon \Gamma \to V$. Then due to (7.6), the map

$$L^2_\pi(\Gamma; V) \to L^2_\pi\big(\Gamma; V^*\big), \quad w \mapsto [y \mapsto A(y)w(y)], \tag{7.10}$$

is well-defined and continuous with norm at most $(1+\gamma)\|D\|_{V\to V^*}$. We assume also that $f \in L^2_\pi(\Gamma; V^*)$.

In order to construct a basis of $L^2_\pi(\Gamma)$, we assume that π is a product measure. Let

$$\pi = \bigotimes_{m=1}^{\infty} \pi_m \tag{7.11}$$

for probability measures π_m on $([-1, 1], \mathscr{B}([-1, 1]))$; see e.g. [4, Section 9] for a general construction of infinite products of probability measures. To avoid degeneracies, we forbid π_m from being a convex combination of finitely many Dirac measures.

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

$$\beta^m_n P^m_n(\xi) := (\xi - \alpha^m_{n-1}) P^m_{n-1}(\xi) - \beta^m_{n-1} P^m_{n-2}(\xi), \quad n \in \mathbb{N}, \tag{7.12}$$

with

$$\alpha^m_n := \int_{-1}^{1} \xi\, P^m_n(\xi)^2 \,\mathrm{d}\pi_m(\xi) \qquad \text{and} \qquad \beta^m_n := \frac{c^m_{n-1}}{c^m_n}, \tag{7.13}$$

where c^m_n is the leading coefficient of P^m_n, $\beta^m_0 := 1$, and P^m_n is chosen as normalized in $L^2_{\pi_m}([0, 1])$. This basis is unique e.g. if c^m_n is chosen to be positive. We note that $\alpha^m_n = 0$ if π_m is symmetric.

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

$$\Lambda := \Big\{\mu \in \mathbb{N}_0^{\mathbb{N}} \,;\, \#\operatorname{supp}\mu < \infty\Big\}, \tag{7.14}$$

where the support of $\mu \in \mathbb{N}_0^{\mathbb{N}}$ is defined as $\operatorname{supp}\mu := \{m \in \mathbb{N} \,;\, \mu_m \neq 0\}$.

The countably infinite tensor product polynomials

$$\mathbf{P} := \big(P_\mu\big)_{\mu\in\Lambda}, \quad P_\mu := \bigotimes_{m=1}^{\infty} P^m_{\mu_m}, \quad \mu \in \Lambda, \tag{7.15}$$

form an orthonormal basis of $L^2_\pi(\Gamma)$, see e.g. [25, Theorem 2.8]. Note that each of these functions depends on only finitely many dimensions,

$$P_\mu(y) = \prod_{m=1}^{\infty} P^m_{\mu_m}(y_m) = \prod_{m\in\operatorname{supp}\mu} P^m_{\mu_m}(y_m), \quad \mu \in \Lambda, \tag{7.16}$$

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

To the basis $\mathbf{P}$ of $L^2_\pi(\Gamma)$, we add a Riesz basis $\Phi = (\varphi_\iota)_{\iota\in\Xi}$ of V, and discretize (7.2) with respect to the product basis $\mathbf{P}\times\Phi = (P_\mu\otimes\varphi_\iota)_{(\mu,\iota)\in\Lambda\times\Xi}$ as

$$\mathbf{A}\mathbf{u} = \mathbf{f} \tag{7.17}$$

with $\mathbf{A} \in \mathscr{L}(\ell^2(\Lambda\times\Xi))$, $\mathbf{f} = (\int_\Gamma \langle f(y),\varphi_\iota\rangle\, P_\mu(y)\,\mathrm{d}\pi(y))_{(\mu,\iota)\in\Lambda\times\Xi} \in \ell^2(\Lambda\times\Xi)$ and $\mathbf{u} = (u_{\mu,\iota})_{(\mu,\iota)\in\Lambda\times\Xi} \in \ell^2(\Lambda\times\Xi)$ representing u through

$$u(y) = \sum_{(\mu,\iota)\in\Lambda\times\Xi} u_{\mu,\iota} P_\mu(y)\varphi_\iota \tag{7.18}$$

with convergence in $L^2_\pi(\Gamma; V)$.

Due to the recursion (7.12),

$$y_m P_\mu(y) = \beta^m_{\mu_m+1} P_{\mu+\varepsilon_m}(y) + \alpha^m_{\mu_m} P_\mu(y) + \beta^m_{\mu_m} P_{\mu-\varepsilon_m}(y) \tag{7.19}$$

for any $\mu \in \Lambda$ and $m \in \mathbb{N}$, where $\varepsilon_m = (\delta_{mn})_{n\in\mathbb{N}}$ is the Kronecker sequence and $P_\nu = 0$ if any $\nu_m < 0$. Consequently, the map

$$\mathbf{K}_m \colon (c_\mu)_{\mu\in\Lambda} \mapsto \big(\beta^m_{\mu_m+1} c_{\mu+\varepsilon_m} + \alpha^m_{\mu_m} c_\mu + \beta^m_{\mu_m} c_{\mu-\varepsilon_m}\big)_{\mu\in\Lambda} \tag{7.20}$$

represents $w(y) \mapsto y_m w(y)$ with respect to the basis $\mathbf{P}$. By [25, Lem. 2.11], $\mathbf{K}_m$ is symmetric and $\|\mathbf{K}_m\|_{\ell^2(\Lambda)\to\ell^2(\Lambda)} \le 1$.

Let $\mathbf{D}, \mathbf{R}_m \in \mathscr{L}(\ell^2(\Xi))$ denote the representations of D and R_m, respectively, in the basis Φ of V, and let $\mathbf{I}$ be the identity on $\ell^2(\Lambda)$. Then

$$\mathbf{A} = \mathbf{I}\otimes\mathbf{D} + \sum_{m=1}^{\infty} \mathbf{K}_m \otimes \mathbf{R}_m \tag{7.21}$$

with convergence in $\mathscr{L}(\ell^2(\Lambda\times\Xi))$, where the tensor products are meant with respect to the usual identification of $\ell^2(\Lambda\times\Xi)$ with $\ell^2(\Lambda)\otimes\ell^2(\Xi)$.

8 Sparse approximations of discrete random operators

8.1 Definition of approximations

Let $(\mathbf{D}_j)_{j\in\mathbb{N}_0}$ and $(\mathbf{R}_{m,j})_{j\in\mathbb{N}_0}$ be approximating sequences of $\mathbf{D}$ and $\mathbf{R}_m$, respectively, such that $\mathbf{D}_j$ is $n_{0,j}$-sparse and $\mathbf{R}_{m,j}$ is $n_{m,j}$-sparse, $m\in\mathbb{N}$. We assume $n_{m,0}=0$ and $n_{m,j}$ is strictly increasing in j for all $m\in\mathbb{N}_0$. Furthermore, let

$$\|\mathbf{D}-\mathbf{D}_j\|_{\ell^2(\Xi)\to\ell^2(\Xi)}\le\bar{e}_{0,j}\quad\text{and}\quad\|\mathbf{R}_m-\mathbf{R}_{m,j}\|_{\ell^2(\Xi)\to\ell^2(\Xi)}\le\bar{e}_{m,j} \tag{8.1}$$

for all $m\in\mathbb{N}$. Such bounds can be computed numerically as in Sect. 6.

For all finitely supported sequences $\mathbf{j}:=(j_m)_{m\in\mathbb{N}_0}$ in $\mathbb{N}_0$, define the operator

$$\mathbf{A}_\mathbf{j}:=\mathbf{I}\otimes\mathbf{D}_{j_0}+\sum_{m=1}^{\infty}\mathbf{K}_m\otimes\mathbf{R}_{m,j_m}. \tag{8.2}$$

Let $\sigma_m:=2$ if the distribution π_m is symmetric, and $\sigma_m:=3$ otherwise. We set $\sigma_0:=1$ and define $\bar{n}_{m,j}:=\sigma_m n_{m,j}$ for $m\in\mathbb{N}_0$. Then for all $j\in\mathbb{N}_0$, $\mathbf{I}\otimes\mathbf{D}_j$ is $\bar{n}_{0,j}$-sparse and $\mathbf{K}_m\otimes\mathbf{R}_{m,j}$ is $\bar{n}_{m,j}$-sparse, $m\in\mathbb{N}$.

Lemma 8.1 *For any finitely supported sequence* $\mathbf{j}=(j_m)_{m\in\mathbb{N}_0}$ *in* $\mathbb{N}_0$, $\mathbf{A}_\mathbf{j}$ *is* $N_\mathbf{j}$*-sparse for*

$$N_\mathbf{j}:=\sum_{m=0}^{\infty}\bar{n}_{m,j_m}, \tag{8.3}$$

and

$$\|\mathbf{A}-\mathbf{A}_\mathbf{j}\|_{\ell^2(\Lambda\times\Xi)\to\ell^2(\Lambda\times\Xi)}\le\sum_{m=0}^{\infty}\bar{e}_{m,j_m}=:\bar{e}_{\mathbf{A},\mathbf{j}}. \tag{8.4}$$

Proof The first part of the assertion follows by construction since $\mathbf{I}$ is 1-sparse and $\mathbf{K}_m$ is σ_m-sparse for all $m\in\mathbb{N}$. Equation (8.4) is a consequence of $\|\mathbf{K}_m\|_{\ell^2(\Lambda)\to\ell^2(\Lambda)}\le 1$ and (7.21).

We use the greedy algorithm from Sect. 3 to select specific $\mathbf{j}$ in (8.2). The cost $c_\mathbf{j}$ and objective $\omega_\mathbf{j}$ are given by

$$c_\mathbf{j}:=N_\mathbf{j}=\sum_{m=0}^{\infty}\bar{n}_{m,j_m}\qquad\text{and}\qquad\omega_\mathbf{j}:=-\bar{e}_{\mathbf{A},\mathbf{j}}=\sum_{m=0}^{\infty}-\bar{e}_{m,j_m}. \tag{8.5}$$

We initialize $\mathbf{j}_0:=\mathbf{0}\in\mathbb{N}_0^{\mathbb{N}_0}$ and construct $(\mathbf{j}_k)_{k\in\mathbb{N}_0}$ recursively by

$$\mathbf{j}_{k+1}:=\texttt{NextOptInf}[\mathbf{j}_k],\quad k\in\mathbb{N}_0, \tag{8.6}$$

using (8.5). Then

$$\mathbf{A}_k := \mathbf{A}_{\mathbf{j}_k}, \quad k \in \mathbb{N}_0, \tag{8.7}$$

defines a sequence of approximations of $\mathbf{A}$. By Lemma 8.1, $\mathbf{A}_k$ is $N_k := N_{\mathbf{j}_k}$-sparse and its distance to $\mathbf{A}$ is bounded by $\bar{e}_{\mathbf{A},k} := \bar{e}_{\mathbf{A},\mathbf{j}_k}$.

Under mild assumptions, (8.7) defines the optimal N_k-sparse approximation of $\mathbf{A}$ given the bounds (8.1) and the estimates in Lemma 8.1.

Assumption 8.A For all $m \in \mathbb{N}$, $n_{m,0} = 0$ and the $(n_{m,j})_{j\in\mathbb{N}_0}$ is strictly increasing. The sequence $(\bar{e}_{m,0})_{m\in\mathbb{N}}$ is in ℓ^1, and $(\bar{e}_{m,j})_{j\in\mathbb{N}_0}$ is nonincreasing. Furthermore, if $i \geq j$, then

$$\frac{-(\bar{e}_{m,i+1} - \bar{e}_{m,i})}{\bar{n}_{m,i+1} - \bar{n}_{m,i}} \leq \frac{-(\bar{e}_{m,j+1} - \bar{e}_{m,j})}{\bar{n}_{m,j+1} - \bar{n}_{m,j}}, \tag{8.8}$$

and $\bar{n}_{m,1}^{-1}(\bar{e}_{m,1} - \bar{e}_{m,0})$ is nonincreasing in m.

The following corollary follows using Theorem 3.4 since Assumption 8.A implies Assumption 3.A for (8.5).

Corollary 8.2 *For all $k \in \mathbb{N}_0$, $\mathbf{j}_k$ minimizes the error bound $\bar{e}_{\mathbf{A},\mathbf{j}}$ among all finitely supported sequences $\mathbf{j}$ in $\mathbb{N}_0$ with sparsity bound $N_{\mathbf{j}} \leq N_k$. Furthermore, if $\bar{e}_{\mathbf{A},k} \neq 0$, then $\mathbf{j}_k$ minimizes $N_{\mathbf{j}}$ among all $\mathbf{j}$ with $\bar{e}_{\mathbf{A},\mathbf{j}} \leq \bar{e}_{\mathbf{A},k}$.*

8.2 Numerical computation

We consider the complexity of a routine $\texttt{Build}_{\mathbf{A}}$ as in Def. 4.4 for constructing columns of $\mathbf{A}_k$, interpreted as bi-infinite matrices. To this end, we assume that such assembly routines are available for $\mathbf{D}$ and $\mathbf{R}_m$, $m \in \mathbb{N}$. More specifically, the routines

$$\begin{aligned}
\texttt{Build}_0[j,\iota] &\mapsto \left[(\lambda_i)_{i=1}^{n_{0,j}}, (d_i)_{i=1}^{n_{0,j}}\right],\\
\texttt{Build}_m[j,\iota] &\mapsto \left[(\lambda_i)_{i=1}^{n_{m,j}}, (r_i^m)_{i=1}^{n_{m,j}}\right], \quad m \in \mathbb{N},
\end{aligned}$$

construct all nonzero elements of the ι-th column of $\mathbf{D}_j$ and $\mathbf{R}_{m,j}$, respectively, using no more than $b_m n_{m,j}$ arithmetic operations and storage locations for a constant b_m independent of j and ι.

Due to the assumptions on $\texttt{Build}_m$, $m \in \mathbb{N}_0$:

Lemma 8.3 *The number of arithmetic operations and storage locations required by a call of $\texttt{Build}_{\mathbf{A}}[k,(\mu,\iota)]$ is bounded uniformly in k by $N_k + \sum_{m=0}^{\infty} b_m n_{m,j_{k,m}}$.*

Remark 8.4 It is often necessary to construct $\mathbf{j}_k$ before calling $\texttt{Build}_{\mathbf{A}}[k,\cdot]$, for example to determine N_k and $\bar{e}_{\mathbf{A},k}$. In this case, we can assume $\mathbf{j}_k$ to be readily available in $\texttt{Build}_{\mathbf{A}}[k,\cdot]$. Otherwise, NextOptInf from Sect. 3 can be used to compute $\mathbf{j}_k$ in the first call of $\texttt{Build}_{\mathbf{A}}[k,\cdot]$. If this is done directly for an arbitrary $k \in \mathbb{N}_0$, it adds $\mathcal{O}(k\log(k))$ to the complexity of $\texttt{Build}_{\mathbf{A}}[k,\cdot]$ even if $\mathcal{N}$ is realized

$\texttt{Build}_{\mathbf{A}}[k,(\mu,\iota)] \mapsto \left[((\nu_i,\lambda_i))_{i=1}^{N_k},(a_i)_{i=1}^{N_k}\right]$

$\left[(\lambda_i)_{i=1}^{n_{0,j_{k,0}}},(d_i)_{i=1}^{n_{0,j_{k,0}}}\right] \longleftarrow \texttt{Build}_0[j_{k,0},\iota]$
for $i=1,\dots,n_{0,j_{k,0}}$ **do** $\left[(\nu_i,\lambda_i),a_i\right] \longleftarrow \left[(\mu,\lambda_i),d_i\right]$
$n \longleftarrow n_{0,j_{k,0}}$
for $m \in \mathbb{N}\,;\ j_{k,m} \geq 1$ **do**
  $\left[(\lambda_i)_{i=1}^{n_{m,j_{k,m}}},(r_i^m)_{i=1}^{n_{m,j_{k,m}}}\right] \longleftarrow \texttt{Build}_m[j_{k,m},\iota]$
  $t \longleftarrow 0$
  for $i=1,\dots,n_{m,j_{k,m}}$ **do**
    $(\nu_{n+t+1},\lambda_{n+t+1}) \longleftarrow (\mu+\varepsilon_m,\lambda_i)$
    $a_{n+t+1} \longleftarrow \beta^m_{\mu_m+1} r_i^m$
    if $\mu_m \geq 1$ **then**
      $(\nu_{n+t+2},\lambda_{n+t+2}) \longleftarrow (\mu-\varepsilon_m,\lambda_i)$
      $a_{n+t+2} \longleftarrow \beta^m_{\mu_m} r_i^m$
    if $\sigma_m = 3$ **then**
      $(\nu_{n+t+3},\lambda_{n+t+3}) \longleftarrow (\mu,\lambda_i)$
      $a_{n+t+3} \longleftarrow \alpha^m_{\mu_m} r_i^m$
    $t \longleftarrow t+\sigma_m$
  $n \longleftarrow n+\sigma_m n_{m,j_{k,m}}$

by a tree data structure, which may dominate e.g. if $N_k \lesssim k$. However, if $\texttt{Build}_{\mathbf{A}}[k,\cdot]$ is called successively for $k \in \mathbb{N}$ and the values $\mathbf{j}_k$, $\mathscr{N}$ and M are cached, then the cost of `NextOptInf` is negligible even if $\mathscr{N}$ is realized by a simple linked list.

8.3 Adaptive application of discrete random operators

In this section, we analyze the structure of the adaptive multiplication routine $\texttt{Apply}_{\mathbf{A}}$ from Sect. 4.2 for a discretized parametric operator $\mathbf{A}$ and the approximating sequence $(\mathbf{A}_k)$ from Sect. 8.1.

By Assumption 8.A, $(N_k)_{k\in\mathbb{N}}$ is strictly increasing, and $N_0 = 0$ since $\mathbf{j}_0 = \mathbf{0}$. By definition, $(j_{k,m})_{k\in\mathbb{N}_0}$ is nondecreasing for all $m \in \mathbb{N}_0$. Therefore, Assumption 8.A implies that $(\bar{e}_{\mathbf{A},k})_{k\in\mathbb{N}_0}$ is nonincreasing. If $\bar{e}_{m,j} \to 0$ as $j \to \infty$ for all $m \in \mathbb{N}_0$, since $(\bar{e}_{m,0})_{m\in\mathbb{N}_0} \in \ell^1$, Corollary 8.2 implies that $\bar{e}_{\mathbf{A},k} \to 0$ as $k \to \infty$. We note that

$$\eta_k = \frac{\bar{e}_{\mathbf{A},k} - \bar{e}_{\mathbf{A},k+1}}{N_{k+1} - N_k} = \frac{\bar{e}_{m_k,j_{k,m_k}} - \bar{e}_{m_k,j_{k,m_k}+1}}{\bar{n}_{m_k,j_{k,m_k}+1} - \bar{n}_{m_k,j_{k,m_k}}}, \tag{8.9}$$

which is nonincreasing in k by construction of $(\mathbf{j}_k)_{k\in\mathbb{N}_0}$, see Lemma 3.3. Consequently, Assumption 4.A is satisfied under the sole additional requirement that $\bar{e}_{m,j} \to 0$ as $j \to \infty$ for all $m \in \mathbb{N}_0$.

Also, since

$$\frac{\bar{e}_{\mathbf{A},k}}{\bar{e}_{\mathbf{A},k+1}} = \frac{\bar{e}_{\mathbf{A},k}}{\bar{e}_{\mathbf{A},k} + \bar{e}_{m_k,j_{k,m_k}+1} - \bar{e}_{m_k,j_{k,m_k}}} \leq \frac{\bar{e}_{m_k,j_{k,m_k}}}{\bar{e}_{m_k,j_{k,m_k}+1}},$$

Assumption 5.A is satisfied if

$$\sup_{m\in\mathbb{N}_0}\sup_{j\in\mathbb{N}_0}\frac{\bar{e}_{m,j}}{\bar{e}_{m,j+1}}<\infty. \tag{8.10}$$

Assuming the sequences $(\mathbf{j}_k)$ and (m_k) are known, the first two parts of the routine $\mathtt{Apply}_{\mathbf{A}}[\mathbf{v},\varepsilon]$ can be used to partition the vector $\mathbf{v}$ into $(\mathbf{v}_{[p]})_{p=1}^{\ell}$ and a negligible remainder term, and to assign to each of these a $k_p\in\mathbb{N}_0$.

The final step of $\mathtt{Apply}_{\mathbf{A}}[\mathbf{v},\varepsilon]$ performs the multiplications

$$\mathbf{z}:=\sum_{p=1}^{\ell}\mathbf{A}_{k_p}\mathbf{v}_{[p]}. \tag{8.11}$$

Using the tensor product structure (7.21), (8.11) can be decomposed into multiplications with the coefficient operators $\mathbf{D}_j$ and $\mathbf{R}_{m,j}$, $m\in\mathbb{N}$.

Let $\mathbf{v}_{[p],\mu}$ denote the μ-th coefficient of $\mathbf{v}_{[p]}$, i.e. $\mathbf{v}_{[p],\mu}=(v_{\mu\iota})_{\iota}$ for $\iota\in\Xi$ such that $(\mu,\iota)\in\Xi_p$. Then assuming π_m is symmetric for all $m\in\mathbb{N}$, $\mathbf{z}=(\mathbf{z}_\mu)_{\mu\in\Lambda}$ with

$$\mathbf{z}_\mu=\sum_{p=1}^{\ell}\left(\mathbf{D}_{j_{k_p,0}}\mathbf{v}_{[p],\mu}+\sum_{m=1}^{M_p}\beta^m_{\mu_m+1}\mathbf{R}_{m,j_{k_p,m}}\mathbf{v}_{[p],\mu+\varepsilon_m}+\beta^m_{\mu_m}\mathbf{R}_{m,j_{k_p,m}}\mathbf{v}_{[p],\mu-\varepsilon_m}\right), \tag{8.12}$$

where $M_p:=\max\{m\in\mathbb{N}_0;\ j_{k_p,m}\neq 0\}$. This does not, however, represent an efficient way to construct $\mathbf{z}$. It is not clear which $\mathbf{z}_\mu$ are nonzero, and many multiplications with $\mathbf{R}_{m,j}$ are repeated. The routine $\mathtt{Multiply}_{\mathbf{A}}$ performs the same computation efficiently, for arbitrary π_m, by looping over p and the support of $\mathbf{v}_{[p]}$.

$\mathtt{Multiply}_{\mathbf{A}}[(\mathbf{v}_{[p]})_{p=1}^{\ell},(k_p)_{p=1}^{\ell}]\mapsto\mathbf{z}$

```
z ⟵ 0
for p = 1, ..., ℓ do
    forall the μ ∈ Λ with v_[p],μ ≠ 0 do
        z_μ ⟵ z_μ + D_{j_{k_p,0}} v_[p],μ
        for m = 1, ..., M_p do
            w ⟵ R_{m,j_{k_p,m}} v_[p],μ
            z_{μ+ε_m} ⟵ z_{μ+ε_m} + β^m_{μ_m+1} w
            if μ_m ≥ 1 then z_{μ−ε_m} ⟵ z_{μ−ε_m} + β^m_{μ_m} w
            if σ_m = 3 then z_μ ⟵ z_μ + α^m_{μ_m} w
```

Remark 8.5 In $\mathtt{Multiply}_{\mathbf{A}}[(\mathbf{v}_{[p]})_{p=1}^{\ell},(k_p)_{p=1}^{\ell}]$, each multiplication with $\mathbf{R}_{m,j}$ is performed only once, and copied to σ_m components of $\mathbf{z}$. This suggests defining $\bar{n}_{m,j}:=n_{m,j}$ for $m\in\mathbb{N}$, without the factor of σ_m from the original definition.

9 s^*-compressibility of discrete random operators

9.1 Preliminary estimates

We assume for the moment that $\mathbf{D}$ and $\mathbf{R}_m$, $m \in \mathbb{N}$, are strictly s-compressible for some $s > 0$. By Proposition 4.3, there is a map $j_0 \colon [0, \infty) \to \mathbb{N}_0$ such that the sparse approximation $\mathbf{D}_{j_0(r)}$ is r-sparse and

$$\|\mathbf{D} - \mathbf{D}_{j_0(r)}\|_{\ell^2(\Xi)\to\ell^2(\Xi)} \leq \bar{e}_{0,j_0(r)} \leq \tilde{d}_{0,s} r^{-s}, \quad r > 0, \tag{9.1}$$

with $\tilde{d}_{0,s} := \tilde{d}_{\mathbf{D},s}$. [2] Similarly, for all $m \in \mathbb{N}$ there is a map $j_m \colon [0, \infty) \to \mathbb{N}_0$ such that the sparse approximation $\mathbf{R}_{m,j_m(r)}$ is $r\sigma_m^{-1}$-sparse and

$$\|\mathbf{R}_m - \mathbf{R}_{m,j_m(r)}\|_{\ell^2(\Xi)\to\ell^2(\Xi)} \leq \bar{e}_{m,j_m(r)} \leq \tilde{d}_{m,s} r^{-s}, \quad r > 0, \tag{9.2}$$

with $\tilde{d}_{m,s} := \sigma_m^s \tilde{d}_{\mathbf{R}_m,s}$.

Lemma 9.1 *If $(\tilde{d}_{m,s})_m \in \ell^{\frac{1}{s+1}}(\mathbb{N}_0)$, then for all $r > 0$ there is a finitely supported sequence $\mathbf{j}(r)$ in $\mathbb{N}_0$ such that $N_{\mathbf{j}(r)} \leq r$ and*

$$\bar{e}_{\mathbf{A},\mathbf{j}(r)} \leq \left(\sum_{m=0}^{\infty} \tilde{d}_{m,s}^{\frac{1}{s+1}} \right)^{s+1} r^{-s}. \tag{9.3}$$

Proof Let $t > 0$ and define $r_m := \tilde{d}_{m,s}^{\frac{1}{s+1}} t$ for all $m \in \mathbb{N}_0$. Set $\mathbf{j} := (j_m(r_m))_{m\in\mathbb{N}_0}$. This sequence is finitely supported since $r_m < 1$ for all but finitely many $m \in \mathbb{N}_0$. By Lemma 8.1,

$$N_{\mathbf{j}} = \sum_{m=0}^{\infty} \bar{n}_{m,j_m(r_m)} \leq \sum_{m=0}^{\infty} r_m = \sum_{m=0}^{\infty} \tilde{d}_{m,s}^{\frac{1}{s+1}} t =: r$$

and

$$\bar{e}_{\mathbf{A},\mathbf{j}} = \sum_{m=0}^{\infty} \bar{e}_{m,j_m(r_m)} \leq \sum_{m=0}^{\infty} \tilde{d}_{m,s} r_m^{-s} = \sum_{m=0}^{\infty} \tilde{d}_{m,s}^{\frac{1}{s+1}} t^{-s} = \left(\sum_{m=0}^{\infty} \tilde{d}_{m,s}^{\frac{1}{s+1}} \right)^{s+1} r^{-s}.$$

If $(\tilde{d}_{m,s})_m$ is not in $\ell^{\frac{1}{s+1}}(\mathbb{N}_0)$, a similar property still holds if we replace the infinite sum by a partial sum. We define the operators

$$\mathbf{A}_{[M]} := \mathbf{I} \otimes \mathbf{D} + \sum_{m=1}^{M} \mathbf{K}_m \otimes \mathbf{R}_m \in \mathscr{L}(\ell^2(\Lambda \times \Xi)). \tag{9.4}$$

[2] Proposition 4.3 initially only implies that the first term in (9.1) is bounded by the third. However, if (9.1) does not hold, we can replace $\bar{e}_{0,j_0(r)}$ by $\tilde{d}_{0,s} r^{-s}$ in (8.1).

Let

$$\|\mathbf{D}\|_{\ell^2(\Xi)\to\ell^2(\Xi)} \le \bar{e}_{0,0} \quad \text{and} \quad \|\mathbf{R}_m\|_{\ell^2(\Xi)\to\ell^2(\Xi)} \le \bar{e}_{m,0}, \quad m \in \mathbb{N}. \tag{9.5}$$

Then by Lemma 8.1,

$$\|\mathbf{A} - \mathbf{A}_{[M]}\|_{\ell^2(\Lambda\times\Xi)\to\ell^2(\Lambda\times\Xi)} \le \sum_{m=M+1}^{\infty} \bar{e}_{m,0}. \tag{9.6}$$

For any $s > 0$, if either

$$\bar{e}_{m,0} \le s\delta_{\mathbf{A},s}(m+1)^{-s-1} \qquad \forall m \in \mathbb{N} \tag{9.7}$$

or

$$\left(\sum_{m=1}^{\infty} \bar{e}_{m,0}^{\frac{1}{s+1}}\right)^{s+1} \le \delta_{\mathbf{A},s}, \tag{9.8}$$

then it follows as in [27, Prop. 4.4] that

$$\sum_{m=M+1}^{\infty} \bar{e}_{m,0} \le \delta_{\mathbf{A},s}(M+1)^{-s} \qquad \forall M \in \mathbb{N}_0. \tag{9.9}$$

We define

$$\bar{e}_{\mathbf{A}_{[M]},\mathbf{j}} := \sum_{m=0}^{M} \bar{e}_{m,j_m}. \tag{9.10}$$

Then for all sequences $\mathbf{j}$ in $\mathbb{N}_0$ with support in $\{0, 1, \dots, M\}$,

$$\bar{e}_{\mathbf{A},\mathbf{j}} = \bar{e}_{\mathbf{A}_{[M]},\mathbf{j}} + \sum_{m=M+1}^{\infty} \bar{e}_{m,0}. \tag{9.11}$$

The following statement is shown analogously to Lemma 9.1.

Lemma 9.2 *For all $M \in \mathbb{N}_0$ and all $r > 0$, there is a sequence $\mathbf{j}(r)$ in $\mathbb{N}_0$ with support in $\{0, 1, \dots, M\}$ such that $N_{\mathbf{j}(r)} \le r$ and*

$$\bar{e}_{\mathbf{A}_{[M]},\mathbf{j}(r)} \le \left(\sum_{m=0}^{M} \tilde{d}_{m,s}^{\frac{1}{s+1}}\right)^{s+1} r^{-s}. \tag{9.12}$$

Proposition 9.3 *Let (9.7) or (9.8) be satisfied for an $s_\sigma > 0$ and*

$$\left(\sum_{m=0}^{M} \tilde{d}_{m,s}^{\frac{1}{s+1}}\right)^{s+1} \le \hat{d}_s M^{t_s}, \quad M \in \mathbb{N}, \tag{9.13}$$

with $\hat{d}_s > 0$ and $t_s \geq 0$. Then for all $r \in [1, \infty)$ there is a finitely supported sequence $\mathbf{j}(r)$ in $\mathbb{N}_0$ such that $N_{\mathbf{j}(r)} \leq r$ and

$$\bar{e}_{\mathbf{A},\mathbf{j}(r)} \leq \left(\hat{d}_s + \delta_{\mathbf{A},s_\sigma}\right) r^{\frac{-s}{1+t_s/s_\sigma}}. \tag{9.14}$$

Proof Let $r \in [1, \infty)$ and $M := \lfloor r^{\frac{s}{s_\sigma + t_s}} \rfloor$. Then for the sequence $\mathbf{j}(r)$ from Lemma 9.2, $\bar{e}_{\mathbf{A}[M],\mathbf{j}(r)} \leq \hat{d}_s M^{t_s} r^{-s} \leq \hat{d}_s r^{\frac{-ss_\sigma}{s_\sigma + t_s}}$. Equation (9.9) implies

$$\sum_{m=M+1}^{\infty} \bar{e}_{m,0} \leq \delta_{\mathbf{A},s_\sigma} (M+1)^{-s_\sigma} \leq \delta_{\mathbf{A},s_\sigma} r^{\frac{-ss_\sigma}{s_\sigma + t_s}},$$

and the assertion follows using (9.11).

9.2 s^*-compressibility

The above estimates combined with Corollary 8.2 show s^*-compressibility of $\mathbf{A}$ with the approximating sequence $(\mathbf{A}_k)_{k\in\mathbb{N}}$ from Sect. 8.1. Define the constants

$$\tilde{c}_m := \max\left(\bar{n}_{m,1}, \sup_{j\in\mathbb{N}} \frac{\bar{n}_{m,j+1}}{\bar{n}_{m,j}}\right) < \infty, \quad m \in \mathbb{N}_0. \tag{9.15}$$

Note that $c_{\mathbf{D}} \leq \tilde{c}_0$ and $c_{\mathbf{R}_m} \leq \sigma_m \tilde{c}_m$ for $m \in \mathbb{N}$.

Theorem 9.4 *Let s^*_δ, $s^*_\sigma \in (0, \infty]$ and assume*

$$\tilde{c} := \sup_{m\in\mathbb{N}_0} \tilde{c}_m < \infty. \tag{9.16}$$

1. *If $(\tilde{d}_{m,s})_m \in \ell^{\frac{1}{s+1}}(\mathbb{N}_0)$ for all $s \in (0, s^*_\delta)$, then $\mathbf{A}$ is s^*-compressible for $s^* = s^*_\delta$.*
2. *If (9.7) or (9.8) holds for all $s \in (0, s^*_\sigma)$ and (9.13) holds for all $s \in (0, s^*_\delta)$ with $t_s \leq \hat{t} < \infty$, then $\mathbf{A}$ is s^*-compressible for*

$$s^* = \frac{s^*_\delta}{1 + \hat{t}/s^*_\sigma}. \tag{9.17}$$

In both cases, $(\mathbf{A}_k)_{k\in\mathbb{N}}$ is a valid approximating sequence with $c_{\mathbf{A}} \leq \tilde{c}$,

$$d_{\mathbf{A},s} \leq \|(\tilde{d}_{m,s})_m\|_{\ell^{\frac{1}{s+1}}(\mathbb{N}_0)}, \quad s \in (0, s^*) \tag{9.18}$$

in the first case and

$$d_{\mathbf{A},s} \leq \inf_{\frac{s\hat{t}}{s^*_\delta - s} < s_\sigma < s^*_\sigma} \left(\hat{d}_{s(1+\hat{t}/s_\sigma)} + \delta_{\mathbf{A},s_\sigma}\right), \quad s \in (0, s^*) \tag{9.19}$$

in the second case.

Proof Condition (9.16) ensures (4.1) for $(\mathbf{A}_k)_{k\in\mathbb{N}}$ since for $k \in \mathbb{N}$ and $j := j_{k,m_k}$, if $j \geq 1$,

$$\frac{N_{k+1}}{N_k} = \frac{N_k + \bar{n}_{m_k,j+1} - \bar{n}_{m_k,j}}{N_k} = \frac{n + \bar{n}_{m_k,j+1}}{n + \bar{n}_{m_k,j}} \leq \frac{\bar{n}_{m_k,j+1}}{\bar{n}_{m_k,j}} \leq \tilde{c}_{m_k},$$

where $n = N_k - \bar{n}_{m_k,j} \geq 0$, and if $j = 0$,

$$\frac{N_{k+1}}{N_k} = \frac{N_k + \bar{n}_{m_k,1}}{N_k} \leq \bar{n}_{m_k,1} \leq \tilde{c}_{m_k}.$$

Let $s \in (0, s^*)$. In case 9.4, Corollary 8.2 and Lemma 9.1 with $r = N_k$ imply

$$\bar{e}_{\mathbf{A},k} \leq \bar{e}_{\mathbf{A},\mathbf{j}(N_k)} \leq \left(\sum_{m=0}^{\infty} \tilde{d}_{m,s}^{\frac{1}{s+1}}\right)^{s+1} N_k^{-s}.$$

In case 9.4, select $s_\delta \in (0, s_\delta^*)$ and $s_\sigma \in (0, s_\sigma^*)$ such that

$$s = \frac{s_\delta}{1 + \hat{t}/s_\sigma}.$$

This is possible since the right hand side is increasing in s_δ and s_σ. By monotonicity, (9.13) holds with $t_s = \hat{t}$. Then Corollary 8.2 and Proposition 9.3 with $r = N_k$ imply

$$\bar{e}_{\mathbf{A},k} \leq \bar{e}_{\mathbf{A},\mathbf{j}(N_k)} \leq \left(\hat{d}_{s_\delta} + \delta_{\mathbf{A},s_\sigma}\right) N_k^{-s}.$$

Equation (9.19) follows since $s_\delta = s(1 + \hat{t}/s_\sigma)$.

9.3 s^*-computability

Under the assumption that the sequence $(\mathbf{j}_k)_{k\in\mathbb{N}_0}$ is available, s^*-computability of $\mathbf{A}$ follows from Theorem 9.4 as a corollary.

Corollary 9.5 *In the setting of Theorem 9.4, if*

$$\sup_{m\in\mathbb{N}_0} b_m < \infty \tag{9.20}$$

for b_m from Sect. 8.2 and the sequences $\mathbf{j}_k$ are given as in Remark 3.8, then $\mathbf{A}$ is s^*-computable and $\mathtt{Build}_{\mathbf{A}}$ is a valid assembly routine.

Proof s^*-compressibility follows from Theorem 9.4. By Lemma 8.3, (9.20) and Remark 3.8, the number of arithmetic operations and storage locations required by a call of $\mathtt{Build}_{\mathbf{A}}[k, \cdot]$ is $\mathcal{O}(N_k)$.

If $\mathbf{j}_k$ are not readily available, Proposition 3.6 implies that recursive application of `NextOptInf` can construct $\mathbf{j}_k$ in $\mathcal{O}(k\log(k))$ time. Thus $\mathbf{A}$ is still s^*-computable if $k\log(k) \lesssim N_k$. As discussed in Remark 8.4, the cost of computing $\mathbf{j}_k$ from $\mathbf{j}_{k-1}$ using `NextOptInf` is only $\mathcal{O}(\log(k))$. Therefore, if `NextOptInf` is used to construct $\mathbf{j}_k$ in the first call of $\texttt{Build}_{\mathbf{A}}[k,\cdot]$, then $\texttt{Build}_{\mathbf{A}}[k,\cdot]$ requires $\mathcal{O}(N_k)$ operations provided that $\mathbf{j}_{k-1}$ is known, for example from a previous call of $\texttt{Build}_{\mathbf{A}}[k-1,\cdot]$.

10 An illustrative example

10.1 An elliptic boundary value problem

As a model problem, 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

$$\begin{aligned} -\nabla\cdot(a(x)\nabla u(x)) &= f(x)\,, \quad x \in G,\\ u(x) &= 0\,, \qquad x \in \partial G. \end{aligned} \tag{10.1}$$

We view f as deterministic, but model the coefficient a as a series

$$a(y,x) := \bar{a}(x) + \sum_{m=1}^{\infty} y_m a_m(x), \tag{10.2}$$

with $y_m \in [-1,1]$ for all $m \in \mathbb{N}$. Hence a depends on a parameter $y = (y_m)_{m=1}^\infty$ in $\Gamma = [-1,1]^\infty$.

We define the parametric operator

$$A(y)\colon H_0^1(G) \to H^{-1}(G)\,, \quad w \mapsto -\nabla\cdot(a(y)\nabla w)\,, \tag{10.3}$$

for $y \in \Gamma$. Due to the linear dependence of A on a,

$$A(y) = D + R(y), \quad R(y) := \sum_{m=1}^{\infty} y_m R_m \qquad \forall y \in \Gamma \tag{10.4}$$

with convergence in $\mathscr{L}(H_0^1(G), H^{-1}(G))$, as assumed in (7.3) and (7.8), for

$$\begin{aligned} D&\colon H_0^1(G) \to H^{-1}(G)\,, \quad w \mapsto -\nabla\cdot(\bar{a}\nabla w)\,,\\ R_m&\colon H_0^1(G) \to H^{-1}(G)\,, \quad w \mapsto -\nabla\cdot(a_m\nabla w)\,, \quad m \in \mathbb{N}. \end{aligned}$$

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

$$\operatorname*{ess\,inf}_{x\in G} \bar{a}(x) \geq \delta^{-1}. \tag{10.5}$$

Since $\|R_m\|_{H_0^1(G)\to H^{-1}(G)} \le \|a_m\|_{L^\infty(G)}$, (7.9) follows from

$$\delta \sum_{m=1}^{\infty} \|a_m\|_{L^\infty(G)} \le \gamma < 1. \tag{10.6}$$

This condition can be loosened by defining $\langle D\cdot,\cdot\rangle$ as the inner product of $H_0^1(G)$, in which case the factor δ in (10.6) vanishes, and $\|a_m\|_{L^\infty(G)}$ is replaced by $\|a_m/\bar{a}\|_{L^\infty(G)}$. We refer to e.g. [25,23,31] for further extensions that still ensure (7.5).

10.2 Optimal finite element discretization

Approximation results for the solution u of (10.1) have been shown in [12] for the case that y_m are uniformly distributed. In this setting, the orthogonal polynomials P_n^m from Sect. 7.2 are Legendre polynomials, normalized with respect to the uniform probability measure on $[-1, 1]$.

Let $(V_j)_{j=0}^\infty$ be a nested sequence of finite element spaces in $H_0^1(G)$ with geometrically increasing dimensions $M_j := \dim V_j$, satisfying

$$\inf_{w_j\in V_j} \|w - w_j\|_{H_0^1(G)} \le C M_j^{-t} |w|_Z \qquad \forall w \in Z\,, \tag{10.7}$$

where Z is a subspace $Z \subset H_0^1(G)$ with seminorm $|\cdot|_Z$ and norm $(\|\cdot\|^2_{H_0^1(G)} + |\cdot|_Z^2)^{1/2}$, such as a higher order Sobolev space. We consider approximations to u in which, for some finite set $\Xi \subset \Lambda$, each coefficient u_μ for $\mu \in \Xi$ is approximated in some finite element space $V_\mu := V_{j(\mu)}$, and the remaining u_μ are set to zero.

If $\mathbf{u} \in \ell^p(\Lambda; H_0^1(G))$ for some $p \in (0, 2)$, then Stechkin's lemma [11, Lem. 5.5] implies that if Ξ_N contains the first $N - 1$ indices μ in a decreasing rearrangement of $\|u_\mu\|_{H_0^1(G)}$, the truncation error satisfies

$$\left(\sum_{\mu\in\Lambda\setminus\Xi_N} \|u_\mu\|^2_{H_0^1(G)}\right)^{1/2} \le \|\mathbf{u}\|_{\ell^p(\Lambda;H_0^1(G))} N^{-s}, \quad s = \frac{1}{p} - \frac{1}{2}. \tag{10.8}$$

Following [12], we select spaces V_μ, $\mu \in \Xi_N$, to match this rate. To this end, suppose $\mathbf{u} \in \ell^q(\Lambda; Z)$ for a $q \in [p, \infty]$. Using a Lagrange multiplier to minimize the total dimension $N_{\mathrm{dof}} := \sum_{\mu\in\Xi_N} M_\mu$, with $M_\mu = \dim V_\mu$, under the condition that the total error is equivalent to N^{-s}, leads to a choice of M_μ proportional to $|u_\mu|_Z^{\frac{2}{2t+1}}$. This approximation has a convergence rate of t with respect to N_{dof} if $t \le \frac{1}{q} - \frac{1}{2}$, which coincides with the rate for a single finite element approximation, see (10.7). If $t \ge \frac{1}{q} - \frac{1}{2}$, the resulting approximation rate is

$$s\,\frac{t}{t + \frac{1}{p} - \frac{1}{q}}. \tag{10.9}$$

This is generally less than the semidiscrete approximation rate s, with equality if $q = p$; this last case is considered in [12, Theorem 5.5].

The above summability assumptions are proven in [12] for the case that $|v|_Z = \|\Delta_v\|_{L^2(G)}$. Then $\mathbf{u} \in \ell^p(\Lambda; H_0^1(G))$ if $(a_m) \in \ell^p(\mathbb{N}; L^\infty(G))$, and $\mathbf{u} \in \ell^q(\Lambda; Z)$ holds under the condition $(a_m) \in \ell^q(\mathbb{N}; W^{1,\infty}(G))$. In this setting, t has a maximal value of $1/d$.

Remark 10.1 A similar analysis can be performed if, instead of choosing M_μ by a continuous optimization problem, the finite element spaces are selected to equidistribute the error among all coefficients u_μ, as in the heuristic from [27,24].[3] Due to (10.7), this is achieved for $M_\mu^t \sim |u_\mu|_Z$. The resulting convergence rate with respect to N_{dof} is

$$\frac{2s}{2s+1}\, t \tag{10.10}$$

if $t \leq 1/q$, and coincides with (10.9) if $t \geq 1/q$. Thus the convergence rate reached by the above optimization procedure is attained only if $tq \geq 1$.

10.3 Application of the adaptive stochastic Galerkin method

In Sect. 7.2, D and R_m are discretized by a Riesz basis of $H_0^1(G)$, such as a wavelet basis, leading to operators $\mathbf{D}$ and $\mathbf{R}_m$ on ℓ^2, which can be interpreted as bi-infinite matrices. Although these matrices are generally not sparse, they can be approximated by sparse matrices, and these approximations are pivotal in the efficient adaptive application of the discrete random operator $\mathbf{A}$. We refer to [30] and references therein for constructions of wavelet bases.

It is shown in [34] that for wavelets of order n, i.e. if the dual wavelets have n vanishing moments, $\mathbf{D}$ and $\mathbf{R}_m$ can be s_δ^*-compressible with $s_\delta^* = (n-1)/d$. This is the highest rate of compressibility that adaptive wavelet methods can take advantage of since the order of the wavelets limits the solution of a generic discrete deterministic problem to the space $\mathscr{A}^s$ for $s < s_\delta^*$, see [19,8]. For higher compressibility, the sparsity of the exact solution becomes the limiting factor in the convergence of adaptive wavelet algorithms.

We consider the example $G := (0, 1)$ and

$$a_m(x) := Cm^{-k} \sin(m\pi x), \quad m \in \mathbb{N}, \tag{10.11}$$

with C sufficiently small such that (10.6) holds. Since trigonometric functions often appear in Karhunen–Loève expansions of random fields, see e.g. [5] and references therin, this academic example is quite representative. We note that $(a_m) \in \ell^p(\mathbb{N}; L^\infty(G))$ and $(a_m) \in \ell^q(\mathbb{N}; W^{1,\infty}(G))$ for any $p > 1/k$ and $q > 1/(k-1)$. Thus $\mathbf{u} \in \ell^p(\Lambda; H_0^1(G))$ and $\mathbf{u} \in \ell^q(\Lambda; H^2(G))$ for the same ranges of p and q

[3] This heuristic is actually used to distribute tolerances for a subproblem in [27,24]; it is not clear whether the resulting error in the approximation of u is distributed evenly among all active coefficients.

by [12]. The resulting approximation rates from Sect. 10.2 are 1 for $k \geq 5/2$ and $\frac{1}{2}(k-\frac{1}{2}) \leq 1$ for $k \leq 5/2$.

As mentioned above, it is realistic to assume that the operators $\mathbf{D}$ and $\mathbf{R}_m$, $m \in \mathbb{N}$, are s_δ^*-compressible with $s_\delta^* \geq 1$. In order to derive s^*-compressibility of the discrete stochastic operator $\mathbf{A}$, Theorem 9.4 requires a degree of summability of the compressibility constants of these operators. Entries in the matrix representations of these operators are zero for basis functions with disjoint supports, and they generally also become insignificant if the supports overlap, but the wavelets have sufficiently different length scales. In this example, the latter effect only sets in once the smaller length scale is below $1/m$. Consequently, we are left with $\mathcal{O}(m)$ significant entries in columns of $\mathbf{R}_m$ corresponding to coarse-scale basis functions.

For any $r > 0$, let $e_{m,r}$ denote the error in an r-sparse approximation of $\mathbf{R}_m$. Then the sparsity required to achieve an error of $e_{m,r} \sim m^{-k} e_{1,\rho}$ in the approximation of $\mathbf{R}_m$ is $r \sim \rho m$. This implies

$$\tilde{d}_{m,s} \sim \sup_{r>0} r^s e_{m,r} \sim \sup_{\rho>0} \rho^s m^s m^{-k} e_{1,\rho} = m^{-(k-s)} \tilde{d}_{1,s}. \tag{10.12}$$

In this setting, the condition $(\tilde{d}_{m,s})_m \in \ell^{\frac{1}{s+1}}(\mathbb{N}_0)$ of Theorem 9.4 is equivalent to $k-s > s+1$, i.e. $s < (k-1)/2$. Hence we can realistically expect s^*-compressibility of $\mathbf{A}$ for $s^* = (k-1)/2$, provided $s_\delta^* \geq s^*$.

For $k \leq 3$, the compression rate s^* is less than or equal to the approximation rate, and thus s^*-compressibility is the limiting factor in the complexity of adaptive wavelet methods for our model problem. For $k \geq 3$, the limited spatial regularity shown in [12] becomes the main obstacle, and the compression rate is larger than the approximation rate given here.

Despite the slightly suboptimal complexity of adaptive wavelet methods due to the compression rate s^* being smaller than the approximation rate, the direct application of these methods to the fully discrete problem improves on the heuristic used in [27,24]. For example, if $k = 3$, then $\mathbf{A}$ is s^*-compressible for $s^* = 1$, and $\mathbf{u} \in \mathscr{A}^s(\Lambda \times \Xi)$ for all $s < 1$. However, if u is approximated by finite elements with the same approximation error in each active coefficient, then the optimal approximation rate is only $5/6$, see Remark 10.1. A similar property holds for any $k \geq (3+\sqrt{5})/2$ since the approximation rate with equidistributed errors is essentially $1 - \frac{1}{2k}$ for $k \geq 2$.

Acknowledgments This research was supported in part by the Swiss National Science Foundation grant No. 200021-120290/1. The author expresses his gratitude to Ch. Schwab for his many insightful remarks and suggestions.

References

1. Babuška, I., Tempone, R., Zouraris, G.E.: Galerkin finite element approximations of stochastic elliptic partial differential equations. SIAM J. Numer. Anal. **42**(2), 800–825 (2004). (electronic)
2. Barinka, A.: Fast evaluation tools for adaptive wavelet schemes. Ph.D. thesis, RWTH Aachen (2005)
3. Barinka, A., Dahlke, S., Dahmen, W.: Adaptive application of operators in standard representation. Adv. Comput. Math. **24**(1–4), 5–34 (2006)

4. Bauer, H.: Wahrscheinlichkeitstheorie, 5th edn. de Gruyter Lehrbuch [de Gruyter Textbook]. Walter de Gruyter & Co., Berlin (2002)
5. Bieri, M., Andreev, R., Schwab, C.: Sparse tensor discretization of elliptic SPDEs. SIAM J. Sci. Comput. **31**(6), 4281–4304 (2009)
6. Bieri, M., Schwab, C.: Sparse high order FEM for elliptic sPDEs. Comput. Methods Appl. Mech. Eng. **198**(13–14), 1149–1170 (2009)
7. Cioica, P., Dahlke, S., Döhring, N., Kinzel, S., Lindner, F., Raasch, T., Ritter, K., Schilling, R.: Adaptive wavelet methods for elliptic stochastic partial differential equations. Tech. rep., DFG 1324 (2011)
8. Cohen, A.: Numerical Analysis of Wavelet Methods, Studies in Mathematics and its Applications, vol. 32. North-Holland Publishing Co., Amsterdam (2003)
9. Cohen, A., Dahmen, W., DeVore, R.: Adaptive wavelet methods for elliptic operator equations: convergence rates. Math. Comput. **70**(233), 27–75 (2001). (electronic)
10. Cohen, A., Dahmen, W., DeVore, R.: Adaptive wavelet methods. II. Beyond the elliptic case. Found. Comput. Math. **2**(3), 203–245 (2002)
11. Cohen, A., DeVore, R., Schwab, C.: Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. Found. Comput. Math. **10**(6), 615–646 (2010)
12. Cohen, A., DeVore, R., Schwab, C.: Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. Anal. Appl. (Singap.) **9**(1), 11–47 (2011)
13. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 3rd edn. MIT Press, Cambridge (2009)
14. Dahlke, S., Fornasier, M., Primbs, M., Raasch, T., Werner, M.: Nonlinear and adaptive frame approximation schemes for elliptic PDEs: theory and numerical experiments. Numer. Methods Partial Differ. Equ. **25**(6), 1366–1401 (2009)
15. Dahlke, S., Fornasier, M., Raasch, T.: Adaptive frame methods for elliptic operator equations. Adv. Comput. Math. **27**(1), 27–63 (2007)
16. Dahlke, S., Raasch, T., Werner, M., Fornasier, M., Stevenson, R.: Adaptive frame methods for elliptic operator equations: the steepest descent approach. IMA J. Numer. Anal. **27**(4), 717–740 (2007)
17. Dahmen, W., Rohwedder, T., Schneider, R., Zeiser, A.: Adaptive eigenvalue computation: complexity estimates. Numer. Math. **110**(3), 277–312 (2008)
18. Deb, M.K., Babuška, I.M., Oden, J.T.: Solution of stochastic partial differential equations using Galerkin finite element techniques. Comput. Methods Appl. Mech. Eng. **190**(48), 6359–6372 (2001)
19. DeVore, R.A.: Nonlinear approximation. In: Acta Numerica, Vol. 7, pp. 51–150. Cambridge University Press, Cambridge (1998)
20. Dijkema, T.J., Schwab, C., Stevenson, R.: An adaptive wavelet method for solving high-dimensional elliptic PDEs. Constr. Approx. **30**(3), 423–455 (2009)
21. Frauenfelder, P., Schwab, C., Todor, R.A.: Finite elements for elliptic problems with stochastic coefficients. Comput. Methods Appl. Mech. Eng. **194**(2–5), 205–228 (2005)
22. Gantumur, T., Harbrecht, H., Stevenson, R.: An optimal adaptive wavelet method without coarsening of the iterands. Math. Comput. **76**(258), 615–629 (2007). (electronic)
23. Gittelson, C.J.: Adaptive Galerkin methods for parametric and stochastic operator equations, Ph.D. thesis. ETH Zürich, ETH Dissertation No. 19533 (2011)
24. Gittelson, C.J.: Adaptive stochastic Galerkin methods: Beyond the elliptic case, Tech. Rep. 2011–2012, Seminar for Applied Mathematics, ETH Zürich (2011)
25. Gittelson, C.J.: Stochastic Galerkin approximation of operator equations with infinite dimensional noise, Tech. Rep. 2011–10. Seminar for Applied Mathematics, ETH Zürich (2011)
26. Gittelson, C.J.: Uniformly convergent adaptive methods for a class of parametric operator equations. ESAIM. Math. Model. Numer. Anal. **46**, 1485–1508 (2012)
27. Gittelson, C.J.: An adaptive stochastic galerkin method for random elliptic operators. Math. Comput. **82**(283), 1515–1541 (2013)
28. Matthies, H.G., Keese, A.: Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. Comput. Methods Appl. Mech. Eng. **194**(12–16), 1295–1331 (2005)
29. Metselaar, A.: Handling wavelet expansions in numerical methods, Ph.D. thesis. University of Twente (2002)
30. Nguyen, H., Stevenson, R.: Finite element wavelets with improved quantitative properties. J. Comput. Appl. Math. **230**(2), 706–727 (2009)

31. Schwab, C., Gittelson, C.J.: Sparse tensor discretization of high-dimensional parametric and stochastic PDEs. In: Acta Numerica, Vol. 20, pp. 291–467. Cambridge University Press, Cambridge (2011)
32. Schwab, C., Stevenson, R.: Space-time adaptive wavelet methods for parabolic evolution problems. Math. Comput. **78**(267), 1293–1318 (2009)
33. Stevenson, R.: Adaptive solution of operator equations using wavelet frames. SIAM J. Numer. Anal. **41**(3), 1074–1100 (2003). (electronic)
34. Stevenson, R.: On the compressibility of operators in wavelet coordinates. SIAM J. Math. Anal. **35**(5), 1110–1132 (2004). (electronic)
35. Stevenson, R.: Adaptive wavelet methods for solving operator equations: an overview. In: Multiscale, Nonlinear and Adaptive Approximation, pp. 543–597. Springer, Berlin (2009)
36. Todor, R.A., Schwab, C.: Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients. IMA J. Numer. Anal. **27**(2), 232–261 (2007)
37. Wan, X., Karniadakis, G.E.: An adaptive multi-element generalized polynomial chaos method for stochastic differential equations. J. Comput. Phys. **209**(2), 617–642 (2005)
38. Wan, X., Karniadakis, G.E.: Multi-element generalized polynomial chaos for arbitrary probability measures. SIAM J. Sci. Comput. **28**(3), 901–928 (2006). (electronic)
39. Xiu, D.: Fast numerical methods for stochastic computations: a review. Commun. Comput. Phys. **5**(2–4), 242–272 (2009)
40. Xiu, D., Karniadakis, G.E.: The Wiener–Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. **24**(2), 619–644 (2002). (electronic)