Institut für Numerische Simulation
Rheinische Friedrich-Wilhelms-Universität Bonn
Wegelerstraße 6 • 53115 Bonn • Germany
phone +49 228 73-3427 • fax +49 228 73-7527
www.ins.uni-bonn.de

M. Griebel and P. Oswald

Stochastic Subspace Correction in Hilbert Space

INS Preprint No. 1717

November 2017

Stochastic Subspace Correction in Hilbert Space

Michael Griebel · **Peter Oswald**

Abstract We consider an incremental approximation method for solving variational problems in infinite-dimensional separable Hilbert spaces, where in each step a randomly and independently selected subproblem from an infinite collection of subproblems is solved. We show that convergence rates for the expectation of the squared error can be guaranteed under weaker conditions than previously established in [**9**].

Keywords infinite space splitting · subspace correction · multiplicative Schwarz · block coordinate descent · greedy · randomized · convergence rates · online learning

Mathematics Subject Classification (2000) 65F10 · 65N22 · 49M27

1 Introduction

The fast solution of quadratic minimization problems or, correspondingly, of large linear systems of equations is an important topic in many application areas of numerical simulation. To this end, iterative algorithms play a major role. They can be formalized by means of *subspace correction methods* for solving positive-definite variational problems in a Hilbert space V using Hilbert space splittings, either in the so-called additive or the multiplicative variant, see [**7**, **15**]. A *Hilbert space splitting* is given by a family of auxiliary Hilbert spaces V_ω, $\omega \in \Omega$, together with a family of bounded linear operators $R_\omega : V_\omega \to V$, such that the span of the subspaces $R_\omega V_\omega \subset V$ is dense in V. The index set Ω can, in principle, be arbitrary. On each V_ω, an appropriate auxiliary positive-definite variational problem is defined (for short, we call this the subproblem on V_ω). In applications, the operators R_ω play the role of extension operators and map subproblem solutions into V, while their duals are restriction operators mapping residuals from V to V_ω. Given such a Hilbert space splitting, in each iteration step a subset of subproblems is solved using the current residual, and the current approximation is corrected (either collectively or successively) by the obtained subproblem solutions in an update step. Examples of the resulting algorithms are the well-known Jacobi and Gauss-Seidel and Kaczmarz algorithms from linear algebra and their block-wise variants, but also domain decomposition methods or multigrid and multilevel techniques from scientific computing.

M. Griebel
Institute for Numerical Simulation, Universität Bonn, Wegelerstr. 6, 53115 Bonn, and Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin
Corresponding author, tel.: +49-228-733437, fax: +49-228-737527,
E-mail: griebel@ins.uni-bonn.de

P. Oswald
Institute for Numerical Simulation, Universität Bonn, Wegelerstr. 6, 53115 Bonn,
E-mail: agp.oswald@gmail.com

The most common way of creating a Hilbert space splitting of V is to choose an at most countable number of closed subspaces $V_i \subset V$ with index set $\Omega \subset \mathbb{N}$ and their natural injection operators as R_i such that their algebraic sum is dense in V. This also explains the name subspace correction methods for the associated algorithms. However, there are many situations (e.g., outer approximation schemes such as finite-difference or nonconforming finite element discretizations), where in order to formulate the iterative method one is naturally confronted with Hilbert spaces $V_\omega \not\subset V$ or nontrivial choices for R_ω. The auxiliary Hilbert spaces V_ω can be one-dimensional or, in a block type fashion, they can have arbitrary (finite) dimension as well. This is an attractive feature to improve runtime efficiency on modern computer systems. Moreover, by choosing and comparing different Hilbert space splittings for the same variational problem on V, one can optimize the performance of iterative solvers. All this adds to the flexibility of using Hilbert space splittings to design and analyze iterative methods for variational problems. However, in this paper we focus on the convergence properties for subspace correction methods for an arbitrarily fixed Hilbert space splitting, and refer to Section 2 for details and precise definitions.

For a (finite or infinite) set of subproblems at hand, the question arises in which order the incremental updates should be made and what the convergence behavior of the resulting iterative method will be. Most conventionally, the order is deterministic and a priorily fixed. This is the case for basically all the classical methods, like Gauss-Seidel, domain decomposition or multigrid methods. The order of traversal through the subproblems is prescribed by the method itself. Examples are lexicographical or so-called red-black orderings for systems stemming from finite element or finite difference discretizations and, additionally, level wise traversal orderings in multigrid algorithms. Besides, for the multiplicative variant, where in each iteration step only one subproblem is solved, greedy methods are popular. Here, the subproblem used in an iteration step is identified according to an optimization criterion such that the actual error is reduced as much as possible. This may substantially improve the convergence of the overall algorithm. A detailed analysis of various greedy approximation methods in a Hilbert space with one-dimensional V_ω is given in the seminal book [**12**]. A simple example from linear algebra is the so-called Gauss-Southwell approach, where the next update variable is that with the largest residuum. Usually, in the case of finitely many subproblems, the determination of the optimal next subproblem can be done exactly, but it involves additional costs. In the case of infinitely many V_ω, this is not possible any more, and heuristic choices are employed there in practical methods.

Besides a deterministic or greedy pick, we may also choose the next subproblem in a random fashion according to a probability distribution ρ defined on Ω, see [**8**] and the references cited therein. The analysis of such stochastic iterations has been an active research topic in large-scale convex optimization, see [**6**] for a recent survey, but also in the area of machine learning and compressed sensing. Compared to the greedy approach, the cost for determining the next subproblem is dramatically reduced to the cost of sampling the underlying probability distribution ρ. Moreover, the random pick is also feasible for infinite Ω, i.e., in the case of infinitely many subproblems. But the question is now what the associated convergence rate (in expectation) will be. For finitely many subspaces, the answer is very encouraging [**8**]: Both greedy and stochastic iterations yield the same exponential rates of convergence, although with different constants, and in the latter case almost surely and in expectation only.

In this article, we deal with the case of an infinite number of subspaces for which a first comparison of greedy and stochastic subspace correction methods was carried out in [**9**] for countable Hilbert space splittings with index set $\Omega = \mathbb{N}$. It was shown that the (much more involved and costly) greedy method converges at an algebraic rate for solutions from a certain class $\mathscr{A}_1$ while basically the same convergence rate can be achieved in expectation by a stochastic subspace correction method on a class $\mathscr{A}_\infty^\rho \subset \mathscr{A}_1$ depending on ρ. Details will be given in the next sections.

The aim of this paper is to show that convergence rates for the expectation of the squared error for can be guaranteed under weaker conditions than previously established in [**9**], namely for Hilbert space splittings with arbitrary infinite index set Ω and for solutions from a class A_2

still depending on ρ, where $\mathscr{A}_\infty^\rho \subset A_2 \subset \mathscr{A}_1$. This result reveals some connection to the theory of approximation algorithms in reproducing kernel Hilbert spaces (RKHS), and may also allow for a wider range of applications of incremental, multiplicative subspace correction methods with randomly picked orderings which may have interesting applications in numerical linear algebra, scientific computing, quadratic optimization, machine learning and compressed sensing.

The remainder of this paper is organized as follows: In Section 2 we give basic notation and introduce our multiplicative subspace correction/approximation algorithm with random picking in the case of a fixed Hilbert space splitting with an infinite (possibly uncountable) index set Ω. Moreover, we give in Theorem 1 and Theorem 2 sharp bounds of its error and thus of its convergence rate in expectation for the class A_2. In Section 3 we discuss various examples of our abstract theory. First, we consider the case of a countable index set Ω and discrete probability measures ρ on it. Moreover, in Lemma 1 we also relate our new function class A_2 to the classes $\mathscr{A}_\infty^\rho$ and $\mathscr{A}_1$, previously used in [**9**]. Then, we consider the case of stochastic approximation in reproducing kernel Hilbert spaces (RKHS) and show that our theory can be applied there as well. Next, we study the case of general unit norm dictionaries and approximation with these, and provide in Theorem 3 a version of our main results from Section 2 with simplified proof. Finally, we deal with a collective approximation problem from [**2**] and show how our theory applies. We conclude in Section 4 with some further remarks on our convergence results.

2 Details and Proofs

Throughout this paper, let V be a separable real Hilbert space. For a given continuous and coercive Hermitian form $a(\cdot,\cdot)$ on V and a bounded linear functional F on V, we consider the variational problem of finding the unique element $u \in V$ such that

$$a(u,v) = F(v) \qquad \forall v \in V. \tag{1}$$

Equivalently, (1) can be formulated as quadratic minimization problem in V or as linear operator equation in the dual space of V. In the following, we use $a(\cdot,\cdot)$ as the scalar product on V, and write $\|v\| = a(v,v)^{1/2}$. Formally, solving (1) is then the same as finding the Riesz representer of F in V.

Our aim is to study a particular instance of an incremental subspace correction (or Schwarz iterative) method for solving (1) based on Hilbert space splittings. Let Ω be a fixed index set equipped with a probability measure ρ (compared to [**9**], we also allow for uncountable Ω, see below for an example). Consider a family $\{V_\omega\}_{\omega\in\Omega}$ of separable real Hilbert spaces, each equipped with a scalar product $a_\omega(\cdot,\cdot)$ and norm $\|v_\omega\|_\omega := a_\omega(v_\omega,v_\omega)^{1/2}$. In principle, the V_ω need not be subspaces of V, nor need the scalar products be related to each other. To relate V and $\{V_\omega\}$, we introduce a family of uniformly bounded linear operators $R_\omega : V_\omega \to V$, $\omega \in \Omega$, i.e., there is a positive constant Λ such that

$$\|R_\omega\|_{V_\omega\to V} = \sup_{\|v_\omega\|_\omega=1} \|R_\omega v_\omega\| \le \Lambda < \infty, \qquad \omega \in \Omega. \tag{2}$$

If $V_\omega \subset V$, one can use the natural injections as R_ω in which case $\Lambda = 1$. Moreover, to avoid trivial problems with approximating arbitrary elements in V, we assume here that V is the closure of $\operatorname{span}\{R_\omega V_\omega : \omega \in \Omega\}$, i.e., that any $v \in V$ can be approximated with arbitrary precision by finite linear combinations of images $R_\omega v_\omega$, $v_\omega \in V_\omega$. Finally, we introduce another family of linear operators $T_\omega : V \to V_\omega$ by the solution of auxiliary variational problems in V_ω:

$$a_\omega(T_\omega v, v_\omega) = a(v, R_\omega v_\omega) \qquad \forall\, v_\omega \in V_\omega, \qquad \omega \in \Omega. \tag{3}$$

It is easy to see that $\|T_\omega\|_{V\to V_\omega} \le \Lambda$ as well. Without loss of generality, we can assume that $\operatorname{Ker}(R_\omega) = \{0\}$ for all ω (otherwise replace V_ω by $V_\omega \ominus_\omega \operatorname{Ker}(R_\omega)$).

From now on, we fix the Hilbert space splitting for V given by $\{V_\omega, R_\omega\}$, and consider an algorithm of the form

$$u^{(m+1)} = \alpha_m u^{(m)} + \xi_m R_{\omega_m} r^{(m)}_{\omega_m}, \qquad r^{(m)}_{\omega_m} = T_{\omega_m}(u - u^{(m)}), \quad m = 0,1,\ldots, \quad u^{(0)} = 0, \tag{4}$$

where $\{\omega_m\}$ is a sequence of independent samples from Ω which are identically distributed according to ρ. In other words, (4) represents a one-step iterative method, where, based on the current iterate $u^{(m)}$, in each step a randomly chosen subproblem is solved.

Concerning the relaxation parameters α_m and ξ_m, as in [**1**, **9**] we have opted to set

$$\alpha_m = 1 - (m+2)^{-1}, \qquad m = 0,1,\ldots, \tag{5}$$

and to choose ξ_m such that the error

$$\delta^2_{m+1} := \|u - u^{(m+1)}\|^2$$

is minimized. This gives the explicit formula

$$\xi_m = \mathrm{argmin}_\xi \|u - \alpha_m u^{(m)} - \xi R_{\omega_m} r^{(m)}_{\omega_m}\|^2 = \frac{F(R_{\omega_m} r^{(m)}_{\omega_m}) - \alpha_m a(u^{(m)}, R_{\omega_m} r^{(m)}_{\omega_m})}{a(R_{\omega_m} r^{(m)}_{\omega_m}, R_{\omega_m} r^{(m)}_{\omega_m})}. \tag{6}$$

Since $r^{(m)}_{\omega_m}$ is defined via (1) and (3) by the variational problem

$$a_\omega(r^{(m)}_{\omega_m}, v_{\omega_m}) = a(u - u^{(m)}, R_{\omega_m} v_{\omega_m}) = F(R_{\omega_m} v_{\omega_m}) - a(u^{(m)}, R_{\omega_m} v_{\omega_m}) \qquad \forall\, v_\omega \in V_\omega, \tag{7}$$

we see that (4) can be executed once $u^{(m)}$ and ω_m are available. The above set of relaxation parameters α_m, ξ_m allows us to follow the proof strategy of [**1**, **9**], and to obtain optimal convergence estimates for (4) on certain dense subspaces in V. Other choices for α_m, ξ_m are possible, however, for the most classical situation of an iteration (4) with constant values $\alpha_m = 1$ and $\xi_m = \xi$ only weaker convergence results are known.

To provide estimates for the expected squared error $\mathbb{E}(\delta^2_m)$, we need the notion of Bochner integrals [**4**]. Given any Bochner-measurable V-valued function $\phi : \omega \in \Omega \to \phi_\omega \in V$, its Bochner integral

$$\mathbb{E}_\rho(\phi) := \int_\Omega \phi_\omega \, d\rho_\omega \tag{8}$$

is well-defined with value in V if the scalar integral

$$\mathbb{E}_\rho(\|\phi\|) := \int_\Omega \|\phi_\omega\| \, d\rho_\omega < \infty \tag{9}$$

exists. The Bochner integral is similarly well-defined if V is replaced by a separable Banach space. In the case of a discrete probability measure on a countable index set Ω, measurability of ϕ is not an issue, in other situations, it needs to be checked. For the following, we assume that for any fixed $e \in V$ the function

$$\tilde\psi : \omega \in \Omega \to \tilde\psi_\omega \in R_\omega(V_\omega) \subset V, \qquad \tilde\psi_\omega := \begin{cases} \frac{R_\omega T_\omega e}{\|R_\omega T_\omega e\|}, & R_\omega T_\omega e \neq 0, \\ 0, & R_\omega T_\omega e = 0, \end{cases} \tag{10}$$

is Bochner-measurable.

Next, we introduce the class $A_2 \equiv A_{2,\rho} \subset V$ which will play a central role in the convergence theory for (4). We say that $u \in V$ belongs to A_2 if there exists a Bochner-measurable function $\phi : \omega \to R_\omega v_\omega$ with $v_\omega \in V_\omega$ for all $\omega \in \Omega$ such that the scalar-valued function $\omega \to \|v_\omega\|_\omega$ is also measurable, and

$$u = \mathbb{E}_\rho(\phi) = \int_\Omega R_\omega v_\omega \, d\rho_\omega, \qquad \mathbb{E}_\rho(\|v_\omega\|^2_\omega) = \int_\Omega \|v_\omega\|^2_\omega \, d\rho_\omega < \infty, \tag{11}$$

Define a norm on A_2 by

$$\|u\|_{A_2} := \inf \mathbb{E}_\rho(\|v_\omega\|_\omega^2)^{1/2}, \tag{12}$$

where the infimum is taken with respect to all admissible representations of u in (11). How this class is related to the classes $\mathscr{A}_p^\gamma$ introduced in [**9**] for discrete measures ρ on countable index sets Ω and other classes used in similar context in the literature will be elaborated on in Section 3.

The central result of this note is the following:

Theorem 1 *If (10) holds and if u belongs to the linear space A_2 induced by the condition (11) then, for the incremental approximation algorithm (4), we have*

$$\mathbb{E}(\delta_m^2) \le \frac{(\Lambda\|u\|_{A_2} + \|u\|)^2}{m+1}, \qquad m = 0, 1, \dots. \tag{13}$$

Proof . We start with an analysis of the error reduction in one recursion step, i.e., with an estimate of $\mathbb{E}_\rho(\delta_{m+1}^2|u^{(m)})$. Throughout the proof we use the notation

$$e^{(m)} := u - u^{(m)}, \qquad \bar\alpha_m := 1 - \alpha_m = (m+2)^{-1}, \qquad w := \alpha_m e^{(m)} + \bar\alpha_m u, \qquad \tilde\psi_\omega := \tilde\psi_\omega e^{(m)},$$

see (10) for the definition of $\tilde\psi_\omega e^{(m)}$. Since ξ_m is given by (6) such that $\delta_{m+1}^2 = \|u - u^{(m+1)}\|^2$ is minimized, and $R_{\omega_m} r_{\omega_m}^{(m)} = R_{\omega_m} T_{\omega_m} e^{(m)}$ is a multiple of $\tilde\psi_{\omega_m}$, we thus obtain

$$\begin{aligned}
\delta_{m+1}^2 &= \min_\xi \|\alpha_m(u - u^{(m)}) + \bar\alpha_m u - \xi R_{\omega_m} r_{\omega_m}^{(m)}\|^2 \\
&= \min_\theta \|w - \theta\tilde\psi_{\omega_m}\|^2 = \|w\|^2 - a(w, \tilde\psi_{\omega_m})^2 \\
&= \alpha_m^2(\delta_m^2 - a(e^{(m)}, \tilde\psi_{\omega_m})^2) + 2\alpha_m\bar\alpha_m(a(e^{(m)}, u) - a(e^{(m)}, \tilde\psi_{\omega_m})a(u, \tilde\psi_{\omega_m})) \\
&\qquad + \bar\alpha_m^2(\|u\|^2 - a(u, \tilde\psi_{\omega_m})^2).
\end{aligned}$$

The measurability assumption for $\tilde\psi_\omega$ allows us to take expectations with respect to the choice of ω_m in the above error representation:

$$\begin{aligned}
\mathbb{E}_\rho(\delta_{m+1}^2|u^{(m)}) &= \alpha_m^2(\delta_m^2 - \mathbb{E}_\rho(a(e^{(m)}, \tilde\psi_\omega)^2)) \\
&\quad + 2\alpha_m\bar\alpha_m(a(e^{(m)}, u) - \mathbb{E}_\rho(a(e^{(m)}, \tilde\psi_\omega)a(u, \tilde\psi_\omega))) + \bar\alpha_m^2(\|u\|^2 - \mathbb{E}_\rho(a(u, \tilde\psi_\omega)^2)).
\end{aligned} \tag{14}$$

Compared to the proof of Theorem 1 b) in [**9**], this is an exact formula rather than an upper estimate for $\mathbb{E}_\rho(\delta_{m+1}^2|u^{(m)})$ which contains the additional term $-\alpha_m^2\mathbb{E}_\rho(a(e^{(m)}, \tilde\psi_\omega)^2)$. Together with the following, more careful estimate of the second term in the right-hand side of (14) this will make the difference.

By the definition of $r_\omega^{(m)} = T_\omega e^{(m)}$ via (3) we have

$$\frac{\|R_\omega r_\omega^{(m)}\|}{\|r_\omega^{(m)}\|_\omega} a(e^{(m)}, \tilde\psi_\omega) = a_\omega(r_\omega^{(m)}, \frac{r_\omega^{(m)}}{\|r_\omega^{(m)}\|_\omega}) \ge a_\omega(r_\omega^{(m)}, \frac{v_\omega}{\|v_\omega\|_\omega}) = \frac{a(e^{(m)}, R_\omega v_\omega)}{\|v_\omega\|_\omega} \tag{15}$$

for any $v_\omega \in V_\omega$ and $\omega \in \Omega$. Together with (2) and (11), this implies

$$a(e^{(m)}, u) = \mathbb{E}_\rho(a(e^{(m)}, R_\omega v_\omega)) \le \Lambda\mathbb{E}_\rho(\|v_\omega\|_\omega a(e^{(m)}, \tilde\psi_\omega)).$$

Thus, we can apply the Cauchy-Schwarz inequality to the second term in the right-hand side of (14):

$$\begin{aligned}
&2\alpha_m\bar\alpha_m(a(e^{(m)}, u) - \mathbb{E}_\rho(a(e^{(m)}, \tilde\psi_\omega)\, a(u, \tilde\psi_\omega)) \\
&\quad \le 2\alpha_m\bar\alpha_m\mathbb{E}_\rho(a(e^{(m)}, \tilde\psi_\omega)(\Lambda\|v_\omega\|_\omega - a(u, \tilde\psi_\omega))) \\
&\quad \le 2\alpha_m\bar\alpha_m\mathbb{E}_\rho(a(e^{(m)}, \tilde\psi_\omega)^2)^{1/2}\mathbb{E}_\rho((\Lambda\|v_\omega\|_\omega - a(u, \tilde\psi_\omega))^2)^{1/2} \\
&\quad \le \alpha_m^2\mathbb{E}_\rho(a(e^{(m)}, \tilde\psi_\omega)^2) + \bar\alpha_m^2(\Lambda^2\mathbb{E}_\rho(\|v_\omega\|_\omega^2) - 2\Lambda\mathbb{E}_\rho(\|v_\omega\|_\omega\, a(u, \tilde\psi_\omega)) + \mathbb{E}_\rho(a(u, \tilde\psi_\omega)^2)) \\
&\quad \le \alpha_m^2\mathbb{E}_\rho(a(e^{(m)}, \tilde\psi_\omega)^2) + \bar\alpha_m^2(\Lambda^2\|u\|_{A_2}^2 + 2\Lambda\|u\|_{A_2}\|u\| + \mathbb{E}_\rho(a(u, \tilde\psi_\omega)^2)),
\end{aligned}$$

where we have used that by (10)

$$|\mathbb{E}_\rho(\|v_\omega\|_\omega a(u,\tilde\psi_\omega))| \le \mathbb{E}_\rho(\|v_\omega\|_\omega)\|u\| \le \mathbb{E}_\rho(\|v_\omega\|_\omega^2)^{1/2}\|u\| = \|u\|_{A_2}\|u\|.$$

After substitution into (14) some terms cancel, and we arrive at the estimate

$$\mathbb{E}_\rho(\delta_{m+1}^2|u^{(m)}) \le \alpha_m^2\delta_m^2 + \bar\alpha_m^2(\Lambda\|u\|_{A_2} + \|u\|)^2 \tag{16}$$

for the expectation of the squared error δ_{m+1}^2 conditioned on $u^{(m)}$. Because of the independence assumption, this gives the recursion for the expected error

$$\mathbb{E}(\delta_{m+1}^2) \le \alpha_m^2\mathbb{E}(\delta_m^2) + \bar\alpha_m^2(\Lambda\|u\|_{A_2} + \|u\|)^2, \qquad m = 0,1,\dots, \tag{17}$$

with $\mathbb{E}_\rho(\delta_0^2) = \|u\|^2$ since we set $u^{(0)} = 0$.

The remaining steps are as in [**9**]. Due to the specific choice of α_m, we can rewrite (17) as a recursion

$$b_{m+1} \le \alpha_m b_m + \bar\alpha_m(\Lambda\|u\|_{A_2} + \|u\|)^2, \qquad m = 0,1,\dots, \qquad b_0 = \|u\|^2,$$

for the new sequence $b_m := (m+1)\mathbb{E}(\delta_m^2)$. Since $\alpha_m + \bar\alpha_m = 1$ this implies $b_m \le (\Lambda\|u\|_{A_2} + \|u\|)^2$ uniformly in m which is equivalent to (13), and concludes the proof of Theorem 1. □

As in [**9**], the proof of Theorem 1 can be modified to yield an estimate valid for arbitrary $u \in V$. This results in the following:

Theorem 2 *If the functions defined in (10) are Bochner-measurable then for arbitrary $u \in V$ the algorithm (4) satisfies*

$$\mathbb{E}(\delta_m^2)^{1/2} \le 2\left(\|u-h\| + \frac{((\Lambda\|h\|_{A_2} + \|h\|)^2 + \|u\|^2)^{1/2}}{(m+1)^{1/2}}\right), \qquad m = 0,1,\dots, \tag{18}$$

where $h \in A_2$ is arbitrary. As a consequence, we have $\mathbb{E}(\delta_m^2) \to 0$ for arbitrary $u \in V$.

Proof. To see (18), write

$$a(e^{(m)},u) - a(e^{(m)},\tilde\psi_{\omega_m})\,a(u,\tilde\psi_{\omega_m}) \le a(e^{(m)},h) - a(e^{(m)},\tilde\psi_{\omega_m})\,a(h,\tilde\psi_{\omega_m}) + \|u-h\|\|e^{(m)}\|,$$

and proceed as above for the first term in the right-hand side, using the assumption $h \in A_2$. Instead of (17), this yields

$$\mathbb{E}(\delta_{m+1}^2) \le \alpha_m^2\mathbb{E}(\delta_m^2) + 2\alpha_m\bar\alpha_m\mathbb{E}(\delta_m)\|u-h\| + \bar\alpha_m^2((\Lambda\|h\|_{A_2} + \|h\|)^2 + \|u\|^2), \tag{19}$$

$m = 0,1,\dots$. The rest of the argument leading to (18) is the same as in the proof of [**9**, Theorem 2]. Since $\operatorname{span}\{R_\omega V_\omega : \omega \in \Omega\} \subset A_2$ is dense in V, the convergence in expectation for arbitrary $u \in V$ follows from (18). □

The bounds in Theorem 1 and Theorem 2 carry over to the stochastic version of orthogonal matching pursuit (OMP), where the recursion (4) is replaced by

$$u^{(m+1)} = P_{W_m}u, \qquad r_{\omega_m}^{(m)} = T_{\omega_m}(u - u^{(m)}), \quad m = 0,1,\dots, \quad u^{(0)} = 0, \tag{20}$$

with P_{W_m} denoting the orthogonal projection onto the subspace

$$W_m := \operatorname{span}(\{R_{\omega_0}r_{\omega_0}^{(0)},\dots,R_{\omega_m}r_{\omega_m}^{(m)}\})$$

in V. This is because, for given $u^{(k)}$ and ω_k, $k = 0,\dots,m$, the error of the stochastic OMP algorithm after the update step satisfies

$$\|u - u^{(m+1)}\| = \|u - P_{W_m}u\| \le \|u - \alpha_m u^{(m)} - \xi_m R_{\omega_m}r_{\omega_m}^{(m)}\|$$

for any choice of ξ_m. Consequently, the estimates for one step of (4) can be applied, and we obtain the same recursions for the expectations of the squared error $\mathbb{E}(\|u - P_{W_m}u\|^2)$ of stochastic OMP as in (17) and (19) for our algorithm (4). Thus, the bounds in Theorem 1 and Theorem 2 hold for stochastic OMP as well. In practice, the stochastic OMP (20) is expected to converge slightly faster than our algorithm (4), at the expense of a more costly evaluation of the projections $P_{W_m}u$ in each step.

Finally, note that the class $A_2 \subset V$ delicately depends on the Hilbert space splitting given by $\{V_\omega, R_\omega\}$ and the probability measure ρ. It can be made more explicit in some cases which we outline in the next section.

3 Examples

3.1 Countable Ω and discrete measures

The most common situation in which our results can be made more explicit is the case of a discrete measure ρ on a countable Ω. To allow for a direct comparison with the results in [**9**], set without loss of generality $\Omega = \mathbb{N}$ and denote $\rho_i := \rho(\{i\}) > 0$. Then the measurability assumptions for (10) and (11) are irrelevant. Thus, $u \in A_2$ is, according to (11), equivalent to the existence of $v_i \in V_i$ such that

$$u = \sum_i \rho_i R_i v_i, \qquad \sum_i \rho_i \|v_i\|_i^2 < \infty,$$

where $\|\cdot\|_i$ is the norm in V_i. Moreover,

$$\|u\|_{A_2}^2 = \inf_{u=\sum_i R_i v_i} \sum_i \rho_i \|v_i\|_i^2.$$

In [**9**], for any sequence $\gamma := \{\gamma_i > 0\}$ and $0 < q \le \infty$, classes $\mathscr{A}_q^\gamma$ were introduced by the requirement that $u \in \mathscr{A}_q^\gamma$ if there are $w_i \in V_i$ such that

$$u = \sum_i R_i w_i, \qquad \|\{\gamma_i^{-1}\|w_i\|_i\}\|_{\ell_q} < \infty.$$

The (quasi-)norm on $\mathscr{A}_q^\gamma$ is given by

$$\|u\|_{\mathscr{A}_q^\gamma} = \inf_{u=\sum_i R_i w_i} \|\{\gamma_i^{-1}\|w_i\|_i\}\|_{\ell_q}.$$

In particular, for the sequence $\gamma = \mathbf{1}$ given by $\gamma_i = 1$, we simply use the notation $\mathscr{A}_q = \mathscr{A}_q^{\mathbf{1}}$.

Lemma 1 *For any given $\{V, a\}$ and $\{V_i, a_i\}_{i\in\mathbb{N}}$ and any discrete probability measure ρ on $\mathbb{N}$, the following continuous embbedings hold with norm ≤ 1:*

$$\mathscr{A}_\infty^\rho \subset A_2 = \mathscr{A}_2^{\sqrt{\rho}} \subset \mathscr{A}_1.$$

Proof. Since

$$\|u\|_{\mathscr{A}_2^{\sqrt{\rho}}}^2 = \inf_{u=\sum_i R_i w_i} \sum_i \rho_i^{-1}\|w_i\|_i^2 = \inf_{u=\sum_i \rho_i R_i v_i} \sum_i \rho_i \|v_i\|_i^2 = \|u\|_{A_2}^2,$$

the equality $A_2 = \mathscr{A}_2^{\sqrt{\rho}}$ is obvious. Take any u of the form $u = \sum_i R_i w_i$. The inequalities

$$\sum_i \rho_i^{-1}\|w_i\|_i^2 \le \sum_i \rho_i \sup_i (\rho_i^{-1}\|w_i\|_i)^2 = (\sup_i \rho_i^{-1}\|w_i\|_i)^2$$

and

$$\sum_i \|w_i\|_i = \sum_i \rho_i^{1/2}(\rho^{-1/2}\|w_i\|_i) \le (\sum_i \rho_i)^{1/2}(\sum_i \rho_i^{-1}\|w_i\|_i^2)^{1/2} = (\sum_i \rho_i^{-1}\|w_i\|_i^2)^{1/2}$$

imply the embeddings $\mathscr{A}_\infty^\rho \subset \mathscr{A}_2^{\sqrt{\rho}}$ and $\mathscr{A}_2^{\sqrt{\rho}} \subset \mathscr{A}_1$, respectively. □

In [**9**], the condition $u \in \mathscr{A}_\infty^\rho$ was shown to be sufficient for estimates essentially identical with (17) and (19) to hold, therefore the present paper improves the results from [**9**] (and extends them to uncountable Ω). On the other hand, as shown in [**1**, **9**] the condition $u \in \mathscr{A}_1$ is sufficient for proving convergence rates similar to (17) and (19) for the weak greedy version of our algorithm (4), where the random choice of ω_m is replaced by a residual-based search for a $\omega_m \in \Omega$ such that

$$\|r_{\omega_m}^{(m)}\|_{\omega_m} \geq \beta \sup_{\omega\in\Omega} \|r_\omega^{(m)}\|_\omega. \tag{21}$$

Here, $\beta \in (0,1]$ is a fixed parameter. In other words, for the specific algorithm (4) the greedy rule (21) of picking the ω_m yields the same convergence bound on a larger class of u than any of the stochastic search algorithms. The drawback of greedy algorithms is the cost of implementing (21) which typically requires the computation of residuals $r_\omega^{(m)}$ for many $\omega \in \Omega$.

3.2 Stochastic approximation in RKHS

Another case where the above theory can be substantiated is the approximation of functions in a reproducing kernel Hilbert space from randomly selected point evaluations. The standard setting [**3**, **14**] is as follows: Let Ω be a compact metric space, and let $K : \Omega \times \Omega \to \mathbb{R}$ be a continuous positive-definite kernel. This kernel defines a Hilbert space H_K with scalar product $(\cdot,\cdot)_K$ whose elements are continuous functions $f : \Omega \to \mathbb{R}$ such that

$$(K_\omega, f)_K = f(\omega) \qquad \forall\, f \in H_K \quad \forall\, \omega \in \Omega. \tag{22}$$

Here, $K_\omega \in H_K$ is given by $K_\omega(\eta) = K(\omega,\eta)$, $\eta \in \Omega$. Now, choose $V = H_K$ with the scalar product $a(\cdot,\cdot) = (\cdot,\cdot)_K$ and consider the family of one-dimensional subspaces $V_\omega \subset V$ spanned by K_ω, $\omega \in \Omega$. In particular, $a_\omega(\cdot,\cdot)$ is the restriction of $(\cdot,\cdot)_K$ to V_ω, and R_ω is the natural injection ($\Lambda = 1$). With this, we compute

$$R_\omega T_\omega f = T_\omega f = \frac{(K_\omega, f)_K}{(K_\omega, K_\omega)_K} K_\omega = \frac{f(\omega)}{K(\omega,\omega)} K_\omega,$$

where in the last step we have used the reproducing kernel property (22). Thus, our algorithm (4) turns into an incremental approximation process, requiring in each step the evaluation of

$$e^{(m)}(\omega_m) = f(\omega_m) - u^{(m)}(\omega_m),$$

where ω_m is chosen randomly and independently from Ω according to a certain probability distribution ρ. This scenario is typical in learning theory [**13**], where the samples $(\omega_m, y_m) \in \Omega \times \mathbb{R}$, which are drawn according to an (unknown) joint probability distribution $\tilde\rho$ on $\Omega \times \mathbb{R}$, become incrementally available, and one tries to recover the regression function

$$f(\omega) = \mathbb{E}_{\tilde\rho}(y|\omega).$$

In the "no-noise" case ($\mathbb{E}_{\tilde\rho}((y - f(\omega))^2|\omega) = 0$ a.e. on Ω), we would have $y_m = f(\omega_m)$ almost surely, while the ω_m are independent samples drawn from Ω according to the marginal distribution $\rho = \tilde\rho_\omega$.

To apply our theory, i.e., to obtain rates for the expectation of the squared error from (13) and (18), we need to check (10) and have to examine the condition $u \in A_2$ and the approximability of $u \in H_K$ by elements from A_2, respectively. The measurability assumptions for (10) and (11) follow from the uniform continuity of the kernel which implies the uniform continuity of the function $\omega \to K_\omega$, and the measurability of the function $\omega \to R_\omega v_\omega = c_\omega K_\omega$ for any measurable scalar-valued function $\omega \to c_\omega$. Thus, $u \in A_2$ if

$$u(\eta) = (u, K_\eta)_K = \mathbb{E}_\rho((c_\omega K_\omega, K_\eta)_K) = \int_\Omega c_\omega K(\omega,\eta)\, d\rho_\omega, \qquad \int_\Omega c_\omega^2\, d\rho_\omega < \infty,$$

i.e., if u is in the image of $L_2(d\rho)$ under the action of the integral operator L_K with kernel K given by the formula

$$(L_K f)(\eta) := \int_\Omega K(\omega,\eta) f(\omega)\, d\rho_\omega.$$

It is well known that the operator L_K is also well-defined on $V = H_K$, that it is trace-class positive semi-definite on H_K, and that $A_2 = L_K(L_2(d\rho)) = L_K^{1/2}(H_K)$. Thus, our result recovers rates for the noiseless case analogous to those known in online learning with kernels for similar approximation algorithms [**5**, **10**, **11**], where the spaces defined in terms of the spectral decomposition of L_K often serve as smoothness classes.

3.3 General unit norm dictionaries

As a third, slightly different but also slightly more general example, let us consider the case when, for a given separable Hilbert space $V = H$ with scalar product $a(\cdot,\cdot) = (\cdot,\cdot)$, we choose a Borel measure ρ concentrated on the unit sphere $\Omega = S_H = \{\omega \in H :\ \|\omega\| = 1\}$ of H. Then, we consider the algorithm (4) with the family $V_\omega := \mathrm{span}(\{\omega\})$ of one-dimensional subspaces of H (again, $a_\omega(\cdot,\cdot) = (\cdot,\cdot)$ on V_ω, R_ω are the natural injections, and $\Lambda = 1$). Since any function of the form $\omega \in S_H \to v_\omega = c_\omega \omega$ is Bochner-measurable if the scalar-valued function $\omega \to c_\omega$ is measurable, we have $u \in A_2$ iff

$$u = \int_{S_H} c_\omega \omega d\rho, \qquad \int_{S_H} c_\omega^2\, d\rho_\omega < \infty. \tag{23}$$

In this case, the proof of (13) can be carried out directly, using the covariance operator $L : H \to H$ given by

$$Lv = \mathbb{E}_\rho((v,\omega)\omega) = \int_{S_H} (v,\omega)\omega\, d\rho_\omega, \qquad v \in H. \tag{24}$$

This operator is positive semi-definite and trace-class, i.e., there is a complete orthonormal system of eigenfunctions ψ_k of L for the subspace

$$\tilde{H} := H \ominus \mathrm{Ker}(L)$$

with associated eigenvalues $\mu_k > 0$ satisfying $\sum_k \mu_k = 1$. The powers L^s, $s > 0$, are well defined on H and act as isometries between $\tilde{H}$ and the Hilbert spaces

$$H_L^s = L^s(H) := \{v = \sum_k \mu_k^s c_k \psi_k :\ \|v\|_{H_L^s} := (\sum_k c_k^2)^{1/2} < \infty\}.$$

The latter serve as smoothness spaces and, as we will see, $u \in H_L^{1/2}$ implies an analog of (13). Indeed, since $\omega \in S_H$ we have $\tilde{\psi}_\omega = \omega$ in (10) for any e. Taking into account (24) the counterpart of (14) reads as follows:

$$\begin{aligned}\mathbb{E}_\rho(\delta_{m+1}^2) &= \alpha_m^2(\delta_m^2 - \mathbb{E}_\rho((e^{(m)},\omega)^2)) \\ &\quad + 2\alpha_m\bar{\alpha}_m((e^{(m)},u) - \mathbb{E}_\rho((e^{(m)},\omega)(u,\omega))) + \bar{\alpha}_m^2(\|u\|^2 - \mathbb{E}_\rho((u,\omega)^2)) \\ &= \alpha_m^2(\delta_m^2 - (Le^{(m)},e^{(m)})) + 2\alpha_m\bar{\alpha}_m((e^{(m)},u) - (Le^{(m)},u)) + \bar{\alpha}_m^2(\|u\|^2 - (Lu,u)).\end{aligned}$$

Assuming $u \in H_L^{1/2}$, i.e., $u = L^{1/2}v$ for some $v \in \tilde{H} \subset H$ with $\|u\|_{H_L^{1/2}} = \|v\|$, we estimate the second term in the right-hand side by

$$\begin{aligned}2\alpha_m\bar{\alpha}_m((e^{(m)},u) - (Le^{(m)},u)) &= 2\alpha_m\bar{\alpha}_m(L^{1/2}e^{(m)}, (L^{-1/2} - L^{1/2})u) \\ &\le 2\alpha_m\bar{\alpha}_m\|L^{1/2}e^{(m)}\|\|(L^{-1/2} - L^{1/2})u\| \\ &\le \alpha_m^2(Le^{(m)},e^{(m)}) + \bar{\alpha}_m^2(\|v\|^2 - 2\|u\|^2 + (Lu,u)).\end{aligned}$$

Substitution and cancellation of several terms yields the following analog of (16):

$$\mathbb{E}_\rho(\delta_{m+1}^2) \le \alpha_m^2 \delta_m^2 + \bar{\alpha}_m^2 \|u\|_{H_L^{1/2}}^2.$$

The rest is as in the above proof of Theorem 1. This results in the following estimate with slightly improved constant.

Theorem 3 *In the setting described in this subsection, the algorithm (4) converges in expectation for arbitrary $u \in H_L^{1/2}$:*

$$\mathbb{E}(\delta_m^2) \le \frac{\|u\|_{H_L^{1/2}}^2}{m+1}, \qquad m = 0,1,\ldots. \tag{25}$$

The analog of (18) is

$$\mathbb{E}(\delta_m^2)^{1/2} \le 2(\|u-h\| + \frac{(\|h\|_{H_L^{1/2}}^2 + \|u\|^2)^{1/2}}{(m+1)^{1/2}}), \qquad m = 0,1,\ldots, \tag{26}$$

valid for any $u \in H$ and $h \in H_L^{1/2}$. Convergence in expectation $\mathbb{E}(\delta_m^2) \to 0$ holds for any $u \in \tilde{H}$. Moreover, the classes A_2 and $H_L^{1/2}$ coincide, with equality of norms $\|u\|_{A_2} = \|u\|_{H_L^{1/2}}$ for any $u \in H_L^{1/2}$.

Proof. The estimate (25) was already established, the modification leading to (26) is similar to the one in the proof of Theorem 2: Since

$$\begin{aligned}(e^{(m)},u) - (Le^{(m)},u) &= (e^{(m)},h) - (Le^{(m)},h) + (e^{(m)},(I-L)(u-h)) \\ &\le (e^{(m)},h) - (Le^{(m)},h) + \|e^{(m)}\|\|u-h\|,\end{aligned}$$

we can proceed for the first term as above, with u replaced by $h \in A_2$, to arrive at

$$\mathbb{E}(\delta_{m+1}^2) \le \alpha_m^2 \mathbb{E}(\delta_m^2) + 2\alpha_m \bar{\alpha}_m \mathbb{E}(\delta_m)\|u-h\| + \bar{\alpha}_m^2(\|h\|_{A_2}^2 + \|u\|^2).$$

The last term results from a rough estimate of the collection of all terms with forefactor $\bar{\alpha}_m^2$ remaining after substitution, namely

$$\begin{aligned}\|h\|_{A_2}^2 - 2\|h\|^2 + (Lh,h) + \|u\|^2 - (Lu,u) &= \|h\|_{A_2}^2 + \|u\|^2 - \|h\|^2 - ((I-L)h,h) - (Lu,u) \\ &\le \|h\|_{A_2}^2 + \|u\|^2.\end{aligned}$$

For the rest of the argument, we again refer to the proof of Theorem 2 b) in [**9**].

It remains to check that $A_2 = H_L^{1/2}$. For $u \in A_2$ satisfying (23) we can write

$$\|u\|_{H_L^{1/2}}^2 = \sum_k \frac{(u,\psi_k)^2}{\mu_k} = \sum_k \left(\int_{S_H} c_\omega(\omega,\mu_k^{-1/2}\psi_k)\, d\rho_\omega \right)^2 = \sum_k (c_\omega, f_{k,\omega})_{L_2(d\rho)}^2 \le \|c_\omega\|_{L_2(d\rho)}^2 < \infty.$$

The last step follows because the functions $f_{k,\omega} := (\omega,\mu_k^{-1/2}\psi_k)$ form an orthonormal system in $L_2(d\rho)$:

$$(f_{k,\omega}, f_{l,\omega})_{L_2(d\rho)}^2 = \int_{S_H} \frac{(\omega,\psi_k))(\omega,\psi_l)}{\mu_k^{1/2}\mu_l^{1/2}}\, d\rho_\omega = \frac{(L\psi_k,\psi_l)}{\mu_k^{1/2}\mu_l^{1/2}} = \delta_{kl}.$$

Moreover, for similar reasons any $u \in A_2$ must be orthogonal to $\mathrm{Ker}(L)$, i.e., belongs to $\tilde{H}$ and is thus in the closure in H of the orthonormal system $\{\psi_k\}$ of eigenfunctions of L. Indeed, if $v \in \mathrm{Ker}(L)$ then we have $(\omega,v) = 0$ almost everywhere on Ω since

$$\int_{S_H} (\omega,v)^2\, d\rho_\omega = (Lv,v) = 0.$$

This implies the desired orthogonality

$$(u,v) = \int_{S_H} c_\omega(\omega, v)\, d\rho_\omega = 0,$$

and shows $u \in H_L^{1/2}$ and $\|u\|_{H_L^{1/2}} \le \|u\|_{A_2}$ for all $u \in A_2$.

Now, take $u \in H_L^{1/2}$, i.e.,

$$u = \sum_k c_k \psi_k, \qquad \|u\|^2_{H_L^{1/2}} = \sum_k \mu_k^{-1} c_k^2 < \infty.$$

We will check that (23) holds with $c_\omega = \sum_k \mu_k^{-1/2} c_k f_{k,\omega}$, which immediately implies $u \in A_2$ and the opposite inequality $\|u\|_{A_2} \le \|u\|_{H_L^{1/2}}$. This is done by verifying that the moments (u, ψ_l) coincide for both representations of u: On the one hand, we have $(u, \psi_l) = c_l$, on the other hand, we have

$$\left(\int_{S_H} c_\omega \omega \, d\rho_\omega, \psi_l \right) = \int_{S_H} \left(\sum_k \mu_k^{-1/2} c_k f_{k,\omega} \right) (\omega, \psi_l)\, d\rho_\omega = \sum_k (\mu_l/\mu_k)^{1/2} c_k (f_{k,\omega}, f_{l,\omega})_{L_2(d\rho)} = c_l$$

by the orthonormality of the system $\{f_{k,\omega}\}$ in $L_2(d\rho)$. □

3.4 Collective approximation

To demonstrate the versatility of the abstract scheme developed in Section 2, we consider a problem raised in [**2**]: Given an n-dimensional subspace V_n of a Hilbert space H and a dictionary D of unit norm elements in H (the condition $D \subset S_H$ is silently kept througout this subsection), construct, by incrementally selecting dictionary elements $w_0, w_1, \ldots$, subspaces $W_m = \operatorname{span}\{\omega_0, \ldots, \omega_m\}$ which approximate V_n well, i.e., for which estimates for the approximation quantities

$$\sigma_m = \sup_{v \in V_n : \|v\|=1} \inf_{w \in W_m} \|v - w\|_H = \sup_{v \in V_n : \|v\|=1} \|v - P_{W_m} v\|_H, \qquad m = 0, 1, \ldots$$

hold. The collective OMP algorithm proposed in [**2**] uses greedy selection of $w_m \in D$ based on computations involving the ortho-projections P_{W_m} onto W_m which become more costly for larger m. It comes with a convergence rate for the quantity

$$\varepsilon_m = \left(\sum_{i=1}^n \|\phi_i - P_{W_m} \phi_i\|_H^2 \right)^{1/2} = \|\Phi - P_{W_m} \Phi\|_{H^n}, \qquad m = 0, 1, \ldots,$$

where $\Phi = (\phi_1, \ldots, \phi_n)$ is an arbitrarily fixed given orthonormal basis in V_n. Obviously, ε_m does not depend on the choice of Φ, and is an upper bound for σ_m.

We apply our results and design algorithms avoiding the projections P_{W_m} while still guaranteeing similar convergence rates. To set the scene, let

$$V := H^n = \{\mathbf{u} = (u_1, \ldots, u_n) : \ u_1, \ldots, u_n \in H\}$$

be equipped with the usual scalar product

$$a(\mathbf{u}, \mathbf{v}) := \sum_{i=1}^n (u_i, v_i), \qquad \mathbf{u}, \mathbf{v} \in V.$$

We identify the index set Ω with the dictionary D, and consider the family

$$V_w := \{\mathbf{v}_w = \mathbf{c} w : \ \mathbf{c} \in \mathbb{R}^n\}, \qquad w \in D,$$

of n-dimensional subspaces of V. For R_w, $w \in D$, we take the natural injections, thus $\Lambda = 1$. The problem in V we want to solve is $\mathbf{u} = \Phi$ or, in variational form, to find $\mathbf{u} \in V$ such that

$$a(\mathbf{u}, \mathbf{v}) = a(\Phi, \mathbf{v}) \qquad \forall\, \mathbf{v} \in V.$$

With this, we have

$$R_w T_w \mathbf{v} = T_w \mathbf{v} = a(\mathbf{v}, w) w,$$

where $a(\mathbf{v}, w) := ((v_1, w), \ldots, (v_n, w)) \in \mathbb{R}^n$.

Independently of the method of choosing $w_m \in D = \Omega$ (randomly or greedy), our algorithm (4)

$$\mathbf{u}^{(m+1)} = \alpha_m \mathbf{u}^{(m)} + \xi_m r_{w_m}^{(m)}, \qquad r_{w_m}^{(m)} = T_{w_m} \mathbf{e}^{(m)} = (\Phi - \mathbf{u}^{(m)}, w_m) w_m, \qquad m = 0, 1, \ldots,$$

when started with $\mathbf{u}^{(0)} = \mathbf{0}$, produces a sequence of $\mathbf{u}^{(m)}$ whose components belong to W_{m-1} if $m > 0$. Thus, we have upper estimates

$$\varepsilon_m \le \delta_{m+1} := \|\Phi - \mathbf{u}^{(m+1)}\|, \qquad m = 0, 1, \ldots.$$

If we choose the w_m, $m = 0, 1, \ldots$, randomly and independently according to a Borel measure ρ with support on $D \subset S_H$, then Theorems 1 and 2 are applicable, and they imply rates (in expectation) for $\Phi \in A_2$ and general Φ in terms of its approximability by elements $\mathbf{h} \in A_2$. Moreover, it is easy to see that the proof of Theorem 3 remains valid if the application of the operators L and L^s, respectively, which are defined on H and depend on ρ, is extended componentwise to $V = H^n$. This way, we obtain the estimate

$$\sigma_m^2 \le \varepsilon_m^2 \le \mathbb{E}(\delta_{m+1}^2) \le \frac{\|\Phi\|_{A_2}^2}{m+1}, \qquad m = 0, 1, \ldots, \tag{27}$$

if $\Phi \in A_2 = (H_L^{1/2})^n$ with norm in A_2 defined as

$$\|\mathbf{v}\|_{A_2}^2 = \sum_{i=1}^n \|u_i\|_{H_L^{1/2}}^2.$$

The counterpart of (26) holds, too: If $\Phi \in H^n$ then for arbitrary $\Psi \in A_2$ we have

$$\sigma_m^2 \le \varepsilon_m^2 \le \mathbb{E}(\delta_{m+1}^2)^{1/2} \le 2(\|\Phi - \Psi\| + \frac{(\|\Psi\|_{A_2}^2 + \|\Phi\|^2)^{1/2}}{(m+1)^{1/2}}), \qquad m = 0, 1, \ldots. \tag{28}$$

These estimates for the expected error decay of our randomized algorithm are qualitatively the same as for the more expensive collective OMP algorithm with weak greedy selection of the w_m proposed in [**2**]. However, the class A_2 is a strict subset of the class $\mathscr{A}_1(D)$ appearing in the convergence theory in [**2**], and depends on the choice for ρ.

The weak greedy version of our algorithm was already analyzed in [**9**] by generalizing earlier results from [**1**]. For completeness, we repeat it here in the setting and notation of Section 2.

Define the class A_1 as the set of all $u \in V$ for which a representation of the form

$$u = \sum_j R_{\omega^j} v_{\omega^j}, \qquad \sum_j \|v_{\omega^j}\|_{\omega^j} < \infty, \quad \omega^j \in \Omega, \tag{29}$$

holds, and set

$$\|u\|_{A_1} := \inf_{u = \sum_j R_{\omega^j} v_{\omega^j}} \sum_j \|v_{\omega^j}\|_{\omega^j}.$$

For countable Ω, A_1 coincides with the class $\mathscr{A}_1$ defined before.

Theorem 4 *If $u \in A_1$, the algorithm (4) with ω_m chosen according to the weak greedy rule (21) possesses the error bound*

$$\delta_m^2 \le \frac{2((\Lambda/\beta)^2\|u\|_{A_1}^2 + \|u\|^2)}{m+1}, \qquad m = 0,1,\dots. \tag{30}$$

Proof. The proof is almost identical to that of Theorem 1. Indeed, using (21) in (15), we have

$$\frac{\Lambda}{\beta}a(e^{(m)},\tilde\psi_{\omega_m}) \ge \frac{1}{\beta}a_{\omega_m}(r_{\omega_m}^{(m)}, \frac{r_{\omega_m}^{(m)}}{\|r_{\omega_m}^{(m)}\|_{\omega_m}}) \ge a_\omega(r_\omega^{(m)}, \frac{r_\omega^{(m)}}{\|r_\omega^{(m)}\|_\omega}) \ge \frac{a(e^{(m)},R_\omega v_\omega)}{\|v_\omega\|_\omega},$$

for any $\omega \in \Omega$ (as before $\tilde\psi_{\omega_m}$ is defined in (10) with $e = e^{(m)}$). Thus, representing $u \in A_1$ as in (29), we arrive at

$$a(e^{(m)},u) = \sum_j (a(e^{(m)},R_{\omega^j}v_{\omega^j}) \le \frac{\Lambda a(e^{(m)},\tilde\psi_{\omega_m})}{\beta}\sum_j \|v_{\omega^j}\|_{\omega^j},$$

and, after taking the infimum over all such representations of u, we get

$$a(e^{(m)},u) \le \frac{\Lambda\|u\|_{A_1}}{\beta}a(e^{(m)},\tilde\psi_{\omega_m}).$$

For the corresponding term of the error representation for δ_{m+1}^2, this yields

$$\begin{aligned}
&2\alpha_m\bar\alpha_m(a(e^{(m)},u) - a(e^{(m)},\tilde\psi_{\omega_m})a(u,\tilde\psi_{\omega_m})\\
&\quad\le 2\alpha_m\bar\alpha_m a(e^{(m)},\tilde\psi_{\omega_m})((\Lambda/\beta)\|u\|_{A_1} - a(u,\tilde\psi_{\omega_m}))\\
&\quad\le \alpha_m^2 a(e^{(m)},\tilde\psi_{\omega_m})^2 + \bar\alpha_m^2((\Lambda/\beta)\|u\|_{A_1} - a(u,\tilde\psi_{\omega_m}))^2,
\end{aligned}$$

and after substitution and cancellation of terms we have

$$\begin{aligned}
\delta_{m+1}^2 &\le \alpha_m^2\delta_m^2 + \bar\alpha_m^2(((\Lambda/\beta)\|u\|_{A_1} - a(u,\tilde\psi_{\omega_m}))^2 + \|u\|^2 - a(u,\tilde\psi_{\omega_m})^2)\\
&\le \alpha_m^2\delta_m^2 + 2\bar\alpha_m^2((\Lambda/\beta)^2\|u\|_{A_1}^2 + \|u\|^2).
\end{aligned}$$

The rest is as before. □

4 Concluding remarks

We conclude with three further remarks.

Remark 1. In the generality considered here, the obtained convergence rates for the expectation of the squared error δ_m^2 of the algorithm (4) for $u \in A_2$ cannot be improved without additional assumptions on ρ or u. To see this, consider the case of a discrete measure ρ concentrated on a complete orthonormal system $\{e_j\} \subset S_H$ in a Hilbert space $V = H$ with scalar product $a(\cdot,\cdot) = (\cdot,\cdot)$, and denote $\rho_j = \rho(\{e_j\}) > 0$, $j \in \Omega = \mathbb{N}$. This is within the setting of Section 3.3. Obviously, we have

$$Lv = \sum_j \rho_j(v,e_j)e_j, \qquad v \in H \quad (\psi_j = e_j,\ \mu_j = \rho_j,\ j \in \mathbb{N}),$$

and $\mathrm{Ker}(L) = \{0\}$. In other words, $\tilde H = H$, and

$$H_L^s = \{u = \sum_j \rho_j^s c_j e_j : \quad \|u\|_{H_L^s}^2 := \sum_j c_j^2 < \infty\}, \qquad s \in \mathbb{R}.$$

Following the reasoning in Remark 5 in [**9**], for any algorithm that produces the iterates $u^{(m)}$ as linear combinations of at most m elements e_j drawn randomly and independently according to ρ, we then have the lower estimate

$$\mathbf{E}(\|u-u^{(m)}\|^2) \geq \sum_j (u,e_j)^2(1-\rho_j)^m.$$

We mention as a side note that this lower bound is achieved for the stochastic OMP method (20). The condition $u \in H^r_L$ is for $r>0$ equivalent to $(u,e_j)=\rho_j^r(v,e_j)$, $j\in\Omega$, for some $v\in H$. Thus, the worst case behavior of the expected squared error of any such algorithm for recovering $u\in H^r_L$ is characterized by

$$\varepsilon_{m,r} := \sup_{0\neq u\in H^r_L} \frac{\mathbf{E}(\|u-u^{(m)}\|^2)}{\|u\|^2_{H^r_L}} \geq \sup_{0\neq v\in H} \frac{\sum_j (v,e_j)^2\rho_j^{2r}(1-\rho_j)^m}{\sum_j (v,e_j)^2} = \sup_j \rho_j^{2r}(1-\rho_j)^m.$$

Since the function $f(t)=t^{2r}(1-t)^m$ takes its maximum for $t\in[0,1]$ at $t_0=(m+2r)^{-1}$, we see that in general no rate better than $\mathrm{O}(m^{-2r})$ can be expected on the class H^r_L. If we take $r=1/2$, we see that Theorem 3 provides an optimal result, in the sense that the upper limit of $m^{2r}\varepsilon_{m,r}$ for $m\to\infty$ is finite *and* strictly positive for any ρ.

However, with other assumptions on u or on the spectral properties of L (as it is custom in learning with kernel methods [**5**, **10**]), one may expect better results.

Remark 2. The choice for the parameters α_m and ξ_m in the algorithm (4) is appropriate if the evaluations in (4) and (6) (in particular, the functional evaluation $F(u^{(m)})$) are exact. If one attempts to analyze the same algorithm with, e.g., an independent additive noise term ε_m the update formula (4) (in addition to independence, assume $\mathbb{E}(\varepsilon_m)=0$, and $\sigma^2:=\mathbb{E}(\|\varepsilon_m\|^2)=$ const. >0) then, in the formulas for δ^2_{m+1} and subsequently in (17), an additional term σ^2 appears in the right-hand side, i.e.,

$$\mathbf{E}(\delta^2_{m+1}) \leq \alpha_m^2\mathbf{E}(\delta^2_m) + \bar{\alpha}_m^2(\Lambda\|u\|_{A_2}+\|u\|)^2+\sigma^2, \qquad m=0,1,\ldots.$$

Now, the term σ^2 renders any attempt of proving $\mathbf{E}(\delta^2_m)\to 0$ meaningless. At the m-th step of the recursion, an additional term of the order $(m+1)\sigma^2/3$ would appear in the final estimate for $\mathbf{E}(\delta^2_m)$ which is subdominant only for small σ^2 and at the initial stages of the iteration. A crude calculation shows that under these assumptions the best possible bound for the expectation of the squared error is

$$\mathbf{E}(\delta^2_m) \approx \frac{\sigma}{(\Lambda\|u\|_{A_2}+\|u\|)} \quad \text{if } m\approx \frac{\Lambda\|u\|_{A_2}+\|u\|}{\sigma}.$$

This says that, on average, the squared error δ^2_m cannot be approximated better than the standard deviation of the additive noise ε_m relative to the size of u which is unsatisfactory. A possible repair is to give up the minimization requirement for ξ_m, and to execute (4) with some suitably chosen sequence $\xi_m\to 0$. This is well understood for kernel methods in online learning, and represents a significant difference between the noisy and noiseless case.

Remark 3. The right-hand sides in the estimates (18) and (26) in Theorem 2 and Theorem 3 have the form of a K-functional for the pairs (V,A_2) and $(H,H^{1/2}_L)$, respectively. This implies that rates of the form

$$\mathbb{E}(\delta^2_m) = \mathrm{O}(m^{-\theta}), \qquad m\to\infty,$$

with exponent $\theta\in(0,1)$ hold for spaces obtained by real interpolation. E.g., in the setting of Theorem 3, we obtain

$$\mathbb{E}(\delta^2_m) \leq C(m+1)^{-2s}\|u\|^2_{H^s_L}, \qquad m=0,1,\ldots, \tag{31}$$

valid for all $u \in H^s_L$ and $0 < s < 1/2$ with a certain fixed constant C. Indeed, $u \in H^s_L$ can be represented as

$$u = \sum_k c_k \mu_k^s \psi_k, \qquad \|u\|^2_{H^s_L} = \sum_k c_k^2,$$

and setting

$$h = \sum_{k:\,\mu_k \geq (m+1)^{-1}} c_k \mu_k^s \psi_k,$$

we have

$$\begin{aligned}
\|h\|^2_{H^{1/2}_L} &= \sum_{k:\,\mu_k \geq (m+1)^{-1}} c_k^2 \mu_k^{2s-1} \leq (m+1)^{1-2s} \|u\|^2_{H^s_L}, \\
\|u-h\|^2 &= \sum_{k:\,\mu_k < (m+1)^{-1}} c_k^2 \mu_k^{2s} \leq (m+1)^{-2s} \|u\|^2_{H^s_L}, \\
\|u\|^2 &= \sum_k c_k^2 \mu_k^{2s} \leq \|u\|^2_{H^s_L}.
\end{aligned}$$

Thus, after substitution into (26), we get (31) with $C = (2(1+\sqrt{2}))^2 < 24$.

Acknowledgement

M. Griebel was partially supported by the project *EXAHD* of the DFG priority program 1648 *Software for Exascale Computing" (SPPEXA)* and by the Sonderforschungsbereich 1060 *The Mathematics of Emergent Effects* funded by the Deutsche Forschungsgemeinschaft. This paper was written while P. Oswald held a Bonn Research Chair sponsored by the Hausdorff Center for Mathematics at the University of Bonn funded by the Deutsche Forschungsgemeinschaft. He is grateful for this support.

References

1. A. Barron, A. Cohen, W. Dahmen, R. DeVore, Approximation and learning by greedy algorithms, Ann. Statistics 3, (2008), 64–94.
2. P. Binev, A. Cohen, O. Mula, J. Nichols, Greedy algorithms for optimal measurements selection in state estimation using reduced models, 2017. <hal-01638177>.
3. V. Bogachev, Gaussian Measures, AMS, Providence RI, 1998.
4. G. Da Prato, J. Zabczyk, Stochastic Equations in Infinite Dimensions, Cambridge Univ. Press, 1992.
5. A. Dieuleveut, F. Bach, Nonparametric stochastic approximation with large step-sizes, Ann. Statistics 44:4 (2016), 1363–1399.
6. O. Fercoq, P. Richtarik, Optimization in high dimensions via accelerated, parallel, and proximal coordinate descent, SIAM Rev. 58:4, (2016), 739–771.
7. M. Griebel and P. Oswald. On the abstract theory of additive and multiplicative Schwarz algorithms, Numer. Math., 70, (1995), 163–180.
8. M. Griebel, P. Oswald, Greedy and randomized versions of the multiplicative Schwarz method, Linear Algebra Appl. 437(7), (2012), 1596–1610.
9. M. Griebel, P. Oswald, Schwarz iterative methods: Infinite space splittings, Constr. Approx. 44:1 (2016), 121–139.
10. J. Lin, D.-X. Zhou, Learning theory of randomized Kaczmarz algorithm, Journal of Machine Learning Research 16 (2015), 3341–3365.
11. P. Tarrés, Y. Yao, Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence, IEEE Transactions on Information Theory 60:9, 5716–5735.
12. V. Temlyakov, Greedy Approximation, Cambridge University Press, 2012.
13. S. Smale, D.-X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge Univ. Press, 2007.
14. H. Wendland, Scattered Data Approximation, Cambridge Univ. Press, 2010.
15. J. Xu, Iterative methods by space decomposition and subspace correction, SIAM Rev. 34:4 (1992), 581–613.