Institut für Numerische Simulation
Rheinische Friedrich-Wilhelms-Universität Bonn
Friedrich-Hirzebruch-Allee 7 • 53115 Bonn • Germany
phone +49 228 73-69828 • fax +49 228 73-69847
www.ins.uni-bonn.de

M. Griebel, P. Oswald

Convergence analysis of online algorithms for vector-valued kernel regression

INS Preprint No. 2302

September 2023

Convergence analysis of online algorithms for vector-valued kernel regression

Michael Griebel · **Peter Oswald**

Abstract We consider the problem of approximating the regression function from noisy vector-valued data by an online learning algorithm using an appropriate reproducing kernel Hilbert space (RKHS) as prior. In an online algorithm, i.i.d. samples become available one by one by a random process and are successively processed to build approximations to the regression function. We are interested in the asymptotic performance of such online approximation algorithms and show that the expected squared error in the RKHS norm can be bounded by $C^2(m+1)^{-s/(2+s)}$, where m is the current number of processed data, the parameter $0 < s \leq 1$ expresses an additional smoothness assumption on the regression function and the constant C depends on the variance of the input noise, the smoothness of the regression function and further parameters of the algorithm.

Keywords vector-valued kernel regression · online algorithms · convergence rates · reproducing kernel Hilbert spaces

Mathematics Subject Classification (2000) 65D15 · 65F08 · 65F10 · 68W27

1 Introduction

In this paper, we deal with the problem of learning the regression function from noisy vector-valued data using an appropriate RKHS as prior. For the relevant background on the theory of kernel methods, see [**4**, **5**, **12**, **13**, **15**] and specifically [**2**, **3**, **11**] in the vector-valued case. Our emphasis is on obtaining estimates for the expectation of the squared error norm in the RKHS H of approximations to the regression function which are built in an incremental way by so-called online algorithms. The setting we use is as follows: Let be given $N \leq \infty$ samples $(\omega_m, y_m) \in \Omega \times Y$, $m = 0, \ldots, N-1$, of an input-output process $\omega \to y$, which are i.i.d. with respect to a (generally unknown) probability measure μ defined on $\Omega \times Y$. For simplicity, let Ω be a compact metric space, Y a separable Hilbert space, and μ a Borel measure. What we are looking for is a regression function $f_\mu : \Omega \to Y$ which, in some sense, optimally represents the underlying input-output process. We deal with algorithms for least-squares regression which aim at finding approximations to the solution

$$f_\mu(\omega) = \mathbb{E}(y|\omega) \in L^2_\rho(\Omega, Y)$$

M. Griebel
Institute for Numerical Simulation, Universität Bonn, Friedrich-Hirzebruch-Allee 7, 53115 Bonn, Germany, and Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany
Corresponding author, tel.: +49-228-7369829, E-mail: griebel@ins.uni-bonn.de

P. Oswald
Institute for Numerical Simulation, Universität Bonn, Friedrich-Hirzebruch-Allee 7, 53115 Bonn, Germany
E-mail: agp.oswald@gmail.com

of the minimization problem

$$\mathbb{E}(\|f(\omega)-y\|_Y^2) = \int_{\Omega\times Y} \|f(\omega)-y\|_Y^2\, d\mu(\omega,y) \longmapsto \min \tag{1}$$

for $f \in L^2_\rho(\Omega,Y)$ from the samples $(\omega_m, y_m), m = 0,\dots,N-1$, where $\rho(\omega)$ is the marginal probability measure generated by $\mu(\omega,y)$ on Ω. On a theoretical level, for the minimization problem (1) to be meaningful, one needs[1]

$$\mathbb{E}(\|y\|_Y^2) = \int_{\Omega\times Y} \|y\|_Y^2\, d\mu(\omega,y) = \int_\Omega \mathbb{E}(\|y\|_Y^2|\omega)\, d\rho(\omega) < \infty.$$

Since solving the discretized least-squares problem

$$\frac{1}{N}\sum_{m=0}^{N-1} \|f(\omega_m)-y_m\|_Y^2 \longmapsto \min \tag{2}$$

for $f \in L^2_\rho(\Omega,Y)$ is an ill-posed problem which does not make sense without further regularization, it is customary to add a prior assumption $f \in H$, where $H \subset L^2_\rho(\Omega,Y)$ is a set of functions $f:\ \Omega \to Y$ such that point evaluations $\omega \to f(\omega)$ are well-defined. Staying within the Hilbert space setting, candidates for H are vector-valued RKHS which we will introduce in the next section by means of a feature map, i.e. a family of bounded linear operators which map into a separable Hilbert space V. Under standard assumptions, the RKHS H and V are isometric.

Starting from an inital guess $f^{(0)} \in H$, the online algorithms we consider build a sequence of successive approximations $f^{(m)} \in H$, where $f^{(m+1)}$ is a linear combination of the previous approximation $f^{(m)}$ and a term involving the residual $y_m - f^{(m)}(\omega_m)$ with respect to the currently processed sample (ω_m, y_m). More precisely, the update formula can be written in the form

$$f^{(m+1)}(\omega) = \alpha_m(f^{(m)}(\omega) + \mu_m K(\omega,\omega_m)(y_m - f^{(m)}(\omega_m))), \qquad m = 0,1,\dots,N-1, \tag{3}$$

where $K(\omega,\theta): Y \to Y$, $\omega,\theta \in \Omega$, is the operator kernel of the RKHS determined by the feature map. The isometry of H and V allows us to rewrite (3) as iteration in V which is convenient for our subsequent analysis, see Section 2 for the details. Our main result, namely Theorem 1, concerns a sharp estimate for the expected squared error $\mathbb{E}(\|f_\mu - f^{(m)}\|_H^2)$ in the RKHS norm which in this generality seems to be new. It holds under standard assumptions on the feature map, the parameters α_m, μ_m in (3) and the smoothness s of $f_\mu \in H$ measured in a scale of smoothness spaces associated with the underlying covariance operator P_ρ. Moreover, it exhibits the optimal error decay rate $s/(s+2)$. Our approach is an extension of earlier work [**8**] on Schwarz iterative methods in the noiseless case, where $y_m = f_\mu(\omega_m)$.

The remainder of this paper is organized as follows: In Section 2 we introduce vector-valued RKHS, define P_ρ and the associated scale of smoothness spaces $V^s_{P_\rho}$. This sets the stage for the specification of our online learning algorithms in V, allows for their subsequent analysis, and enables us to formulate our main convergence result, namely Theorem 1. In Section 3 we review related results from the literature and compare them to our new result. In Section 4 we then provide the detailed proof of Theorem 1. In Section 5 we give further remarks on Theorem 1, discuss the advantages and limitations of our approach and consider a simple special case of learning an element u of a Hilbert space V from noisy measurements of its coefficients with respect to a complete orthonormal system (CONS) in V.

[1] To obtain quantitative convergence results, stronger conditions on the random variable y such as uniform boundedness μ-a.e. are often imposed in the literature, we will not do this here.

2 Setting and main result

Let us first introduce our approach to vector-valued RKHS H of functions $f: \Omega \to Y$, where Y is a separable Hilbert space. To this end, note that such a RKHS H can be implicitly introduced by a family $\mathbf{R} = \{R_\omega\}_{\omega\in\Omega}$ of bounded linear operators $R_\omega : Y \to V)$, where V is another separable Hilbert space (we will silently assume that V is infinite-dimensional). More precisely, under the condition

$$\bigcap_{\omega\in\Omega} \ker(R^*_\omega) = \{0\}, \tag{4}$$

the space H consists of maps of the form

$$f_v(\omega) := R^*_\omega v, \quad \omega \in \Omega,$$

and

$$\|f_v\|_H := \|v\|_V, \qquad v \in V.$$

Thus, H and V are isometric which allows us to easily switch between H and V in the sequel. In the literature, the map $\omega \to R_\omega$ (or sometimes $\omega \to R^*_\omega$) is called feature map defining the RKHS H.

To simplify our further considerations, we will assume that

$$R_\omega v \in C(\Omega, Y) \qquad \forall\, v \in V. \tag{5}$$

This condition is the continuity of the operator family $\mathbf{R}$ in the strong operator topology and ensures Bochner integrability of functions from Ω into Y and V, respectively, appearing in the formulas below. Due to the assumed compactness of Ω, it also implies

$$\|R_\omega\|^2_{Y\to V} \le \Lambda < \infty, \tag{6}$$

with some $\Lambda < \infty$. Another consequence is that the operator kernel

$$K(\omega,\theta) := R^*_\omega R_\theta : Y \to Y, \qquad \omega,\theta \in \Omega,$$

associated with the vector-valued RKHS H is a Mercer kernel. The condition (6) is equivalent to the uniform boundedness

$$\|K(\omega,\theta)\|_Y \le \Lambda, \qquad \omega,\theta \in \Omega, \tag{7}$$

of the operator kernel K. Moreover, (6) is equivalent to the uniform boundedness of the operator family

$$P_\omega := R_\omega R^*_\omega : V \to V, \qquad \omega \in \Omega,$$

in V, i.e.,

$$\|P_\omega\|_V \le \Lambda, \qquad \omega \in \Omega. \tag{8}$$

For fixed V and $\mathbf{R}$ satisfying the above properties, instead of (1) one now seeks $u \in V$ such that $f_u(\omega) = R^*_\omega u \in H$ is the minimizer of the problem[2]

$$J(v) := \mathbb{E}(\|f_v - y\|^2_Y) = \int_{\Omega\times Y} \|R^*_\omega v - y\|^2_Y \, d\mu(\omega,y) \longmapsto \min. \tag{9}$$

The solution u of this quadratic minimization problem on V, if it exists, must satisfy the necessary condition

$$\mathbb{E}((R^*_\omega u - y, R^*_\omega w)_Y) = \mathbb{E}((P_\omega u - R_\omega y, w)_V) = 0 \qquad \forall\, w \in V.$$

[2] The symbol $\mathbb{E}$ denotes expectations of random variables with respect to the underlying probability space which may vary from formula to formula but should be clear from the context.

This condition is equivalent to the linear operator equation

$$P_\rho u = \mathbb{E}(R_\omega y), \qquad P_\rho := \mathbb{E}(P_\omega) = \mathbb{E}(R_\omega R_\omega^*), \tag{10}$$

in V. The operator $P_\rho : V \to V$ defined in (10), which plays the role of a covariance operator, is bounded and symmetric positive definite. The boundedness of $P_\rho : V \to V$, together with the estimate

$$\|P_\rho\|_V \le \Lambda,$$

follows from (6). The spectrum of P_ρ is contained in $[0,\Lambda]$ and we have $\ker(P_\rho) = \{0\}$ due to (4). Moreover, we will assume that P_ρ is compact. A sufficient condition, which is often satisfied in applications, is the trace class property for P_ρ which in particular holds if the operators R_ω, $\omega \in \Omega$, have uniformly bounded finite rank.

Note here that the assumptions on Ω and $\mathbf{R}$ can be weakened, see for instance [**2**], and that the compactness of P_ρ is only used as technical simplification. In particular, the latter allows us to define the scale of smoothness spaces $V^s_{P_\rho}$, $s \in \mathbb{R}$, generated by P_ρ using the complete orthonormal system (CONS) $\Psi := \{\psi_k\}$ of eigenvectors of P_ρ and associated eigenvalues $\lambda_1 \ge \lambda_2 \ge \ldots > 0$ with limit 0 in V as follows: $V^s_{P_\rho}$ is the completion of $\mathrm{span}(\Psi)$ with respect to the norm

$$\|\sum_k c_k \psi_k\|_{V^s_{P_\rho}} = \left(\sum_k \lambda_k^{-s} c_k^2\right)^{1/2},$$

which is well defined on $\mathrm{span}(\Psi)$ for any s. These spaces will appear in the investigation below.

For our convergence analysis in V, we make another simplifying assumption, namely that

$$f_\mu = f_u, \text{ where } u \in V \text{ is the unique minimizer for (9).} \tag{11}$$

In particular, this means that $\mathbb{E}(R_\omega y) \in \mathrm{ran}(P_\rho)$ and that (10) holds.[3]

For a given prior RKHS H induced by the operator family $\mathbf{R}$ with associated space V and for given samples (ω_m, y_m), $m = 0, \ldots, N-1$, with finite N, the standard regularization of the ill-posed problem (2) is to find the minimizer $u_N \in V$ of the minimization problem

$$J_N(v) := \frac{1}{N}\sum_{m=0}^{N-1} \|R^*_{\omega_m} v - y_m\|_Y^2 + \kappa_N \|v\|_V^2 \longmapsto \min \tag{12}$$

on V, where $\kappa_N > 0$ is a suitable regularization parameter. To compare with (2) or (9), recall that $f_v(\omega_m) = R^*_{\omega_m} v$ are the function values of a function in H. Using the representer theorem for Mercer kernels [**11**], this problem leads to a linear system with a typically dense and ill-conditioned $N \times N$ matrix. There is a huge body of literature, especially in the scalar-valued case $Y = \mathbb{R}$, devoted to setting up, analyzing and solving this problem for fixed N.

We focus here on online learning algorithms for finding approximations to the regression function $f_\mu = f_u$ and are interested in their asymptotic performance, i.e., we assume that N is not fixed (set formally $N = \infty$) and that the i.i.d. samples (ω_m, y_m) become available one by one by a random process, $m = 0, 1, \ldots$. The task of an online algorithm, now viewed as approximation process in V, is then to recover $u \in V$ satisfying (11) from this stream of samples.

To this end, we define the noise term

$$\varepsilon_\omega := y - f_u(\omega) = y - R^*_\omega u, \qquad \omega \in \Omega,$$

[3] If $\mathbb{E}(R_\omega y) \notin \mathrm{ran}(P_\rho)$ then the usual alternative is to study estimates for the squared error $\|f_v - f_W\|^2_{L^2_\rho(\Omega,Y)}$ between $f_v \in H$ (obtained by an approximation process to f_μ from within H) and the $L^2_\rho(\Omega,Y)$ orthoprojection f_W of f_μ onto W, where W is the closure of H in $L^2_\rho(\Omega,Y)$. Since our main result in this section is about norm convergence to u in V or, equivalently, to f_u in H, we will not pursue this option.

which is a Y-valued random variable on $\Omega \times Y$ (to keep the notation short, the depencence on y is not explicitly shown). Since $f_u(\omega) = f_\mu(\omega)$ by (11), we have $\mathbb{E}(\varepsilon_\omega|\omega) = 0$ for any $\omega \in \Omega$. Moreover, the noise variance

$$\sigma^2 := \mathbb{E}(\|\varepsilon_\omega\|_Y^2) = \mathbb{E}(\|y - f_\mu\|_Y^2) \tag{13}$$

with respect to $f_u \in H$ is finite since $\mathbb{E}(\|y\|_Y^2) < \infty$ was assumed in the first place. The value of σ characterizes the average size of the noise $y - f_\mu(\omega)$ on Ω measured in the Y norm.

We consider online algorithms of the standard form

$$u^{(m+1)} = \alpha_m(u^{(m)} + \mu_m R_{\omega_m}(y_m - R^*_{\omega_m} u^{(m)})), \qquad m = 0,1,\ldots, \tag{14}$$

where, at each step, the used sample (ω_m, y_m) is i.i.d. drawn according to the probability measure μ and, consequently, ω_m is i.i.d. drawn according to the marginal probability measure ρ on Ω. Traditionally, the parameters α_m and μ_m are called regularization parameter and step-size parameter (or learning rate), respectively. As to be expected, after applying R^*_ω to both sides in (14) and setting

$$f^{(m)}(\omega) := f_{u^{(m)}}(\omega) = R^*_\omega u^{(m)}, \qquad \omega \in \Omega,$$

we arrive at the online algorithm (3) in H.

The online algorithm (14) is a particular instance of a randomized Schwarz approximation method associated with $\mathbf{R}$. Its noiseless version, where $y_m = R^*_{\omega_m} u$, was studied in [**8**]. Our goal is to derive convergence results for the expected squared error $\mathbb{E}(\|u - u^{(m)}\|_V^2)$, $m = 1,2,\ldots$, which corresponds to convergence estimates in the RKHS H. As to be expected, such estimates will once more require additional smoothness assumptions on u in the form $u \in V^s_{P_\rho}$ with $0 < s \le 1$. However, in contrast to the noiseless case [**8**], they also include a dependence on the noise variance σ^2 in addition to the dependence on the initial error $\|e^{(0)}\|^2$. The prize to pay for convergence is a certain decay of the step-sizes $\mu_m \to 0$ which is typical for stochastic approximation algorithms. More precisely, throughout the remainder of this paper, we set

$$\alpha_m = \frac{m+1}{m+2}, \qquad \mu_m = \frac{A}{(m+1)^t}, \qquad m = 0,1,\ldots, \tag{15}$$

where the parameters $1/2 < t < 1$ and $0 < A \le (2\Lambda)^{-1}$ will be properly fixed later on. In the language of learning algorithms, this is a so-called regularized online algorithm, compared to unregularized online algorithms with $\alpha_m = 1$. Our main result is as follows:

Theorem 1 *Let Y, V be separable Hilbert spaces, Ω be a compact metric space, μ be a Borel probability measure on $\Omega \times Y$, and ρ the marginal Borel probability measure on Ω induced by μ. Assume that*

$$\mathbb{E}(\|y\|_Y^2) = \int_{\Omega \times Y} \|y\|_Y^2 \, d\mu < \infty.$$

For the operator family $\mathbf{R} = \{R_\omega\}_{\omega\in\Omega}$, we require the conditions (4-6). We further assume that the operator $P_\rho = \mathbb{E}(R_\omega R^*_\omega)$ is compact. Finally, we assume (11) and that $u \in V^s_{P_\rho}$ for some $0 < s \le 1$.

Consider the online learning algorithm (14), where $u^{(0)} \in V$ is arbitrary, the parameters α_m, μ_m are given by (15) with $t = t_s := (1+s)/(2+s)$ and $A = 1/(2\Lambda)$ and the random samples (ω_m, y_m), $m = 0,1,\ldots,N \le \infty$, are i.i.d. with respect to μ. Then the expected squared error $\mathbb{E}(\|u - u^{(m)}\|_V^2)$ in V satisfies

$$\mathbb{E}(\|f_\mu - f^{(m)}\|_H^2) = \mathbb{E}(\|u - u^{(m)}\|_V^2) \le C^2 (m+1)^{-s/(2+s)}, \qquad m = 1,2,\ldots,N, \tag{16}$$

where $f^{(m)} = f_{u^{(m)}}$, $C^2 = 2\|e^{(0)}\|_V^2 + 2\|u\|_V^2 + 8\Lambda^s \|u\|^2_{V^s_{P_\rho}} + \sigma^2/\Lambda$ and the noise variance σ^2 is defined in (13).

In this generality, Theorem 1 has not yet appeared in the literature, at least to our knowledge. Its proof is carried out in Section 4. For the parameter range $0 < s \le 1$, the exponent $-s/(2+s)$ in the right-hand side of (16) is best possible under the general conditions stated in Theorem 1. Estimates of the form (16) also hold for arbitrary values $1/2 < t < 1$ and $0 < A \le 1/(2\Lambda)$ admissible in (15), albeit with non-optimal exponents depending on t and different constants C varying with t and A. Note that estimates for

$$\mathbb{E}(\|f_u - f_{u^{(m)}}\|^2_{L^2_\rho(\Omega,Y)}) = \mathbb{E}(\|P_\rho^{1/2} e^{(m)}\|_V^2) = \mathbb{E}((P_\rho e^{(m)}, e^{(m)})_V)$$

with respect to the weaker $L^2_\rho(\Omega,Y)$ norm are of great interest but cannot be obtained within our framework. We will comment on these issues in the concluding Section 5.

There is a huge amount of literature devoted to the convergence theory of various versions of the algorithm (14), especially for the scalar-valued case $Y = \mathbb{R}$. In particular, (14) is often considered in the so-called finite horizon case, where $N < \infty$ is fixed and the step-sizes μ_m are chosen in dependence on N such that expectations such as $\mathbb{E}(\|u - u^{(N)}\|_V^2)$ or $\mathbb{E}(\|f_u - f_{u^{(N)}}\|^2_{L^2_\rho(\Omega,Y)})$, respectively, are optimized for the final approximation $u^{(N)}$. We provide a brief discussion of known results in the next section.

3 Results related to Theorem 1

Given the vast number of publications on convergence rates for learning algorithms, we will only present a selection of results concentrating on the RKHS setting and online algorithms similar to (14). The results we cite are often stated and proved for the scalar-valued case $Y = \mathbb{R}$, even though some authors claim that their methods extend to the case of an arbitrary separable Hilbert space Y with minor changes. One of the first papers on the vector-valued case is [**1**], where the authors provide upper bounds in probability for the $L^2_\rho(\Omega,Y)$ error of f_{u_N} if $N \to \infty$ and $\kappa_N \to 0$, where u_N is the solution of (12). These bounds depend in a specific way on the smoothness of $u \in V^s_{P_\rho}$, $0 \le s \le 1$, and on the spectral properties of P_ρ. Note that in [**1**] and in many other papers stronger assumptions on the compactness of P_ρ compared to our assumptions are made and that bounds in probability do not automatically imply bounds in expectation. Moreover, the error measured in the $L^2_\rho(\Omega,Y)$ norm is with respect to f_{u_N} and not with respect to approximations such as $f_{u^{(m)}}$, $m \le N$, which are produced by a particular algorithm comparable with (14).

In [**14**], the authors provide estimates in probability for an algorithm similar to (14) for the scalar-valued case $Y = \mathbb{R}$. They treat both, convergence in $L^2_\rho(\Omega,\mathbb{R})$ and H norms. There, the main additional assumption needed for the application of certain results from martingale theory is that, for some constant $M_\rho < \infty$, the random variable y satisfies

$$|y| \le M_\rho$$

a.e. on the support of ρ. If $u^{(0)} = 0$ (as assumed in [**14**]) then this assumption implies bounds for $\|e^{(0)}\|_V = \|u\|_V$ and σ with constants depending on M_ρ. Up to the specification of constants and using the notation of the present paper, the convergence result for the H norm stated in [**14**, Theorem B] reads as follows: Consider the online algorithm (14) with starting value $u^{(0)} = 0$ and parameters

$$\alpha_m = \frac{m+m_0-1}{m+m_0}, \qquad \alpha_m\mu_m = \frac{A}{(m+m_0)^{(s+1)/(s+2)}}, \qquad m = 0,1,\ldots,$$

for some (large enough) m_0 and suitable A. Then, if $u \in V^s_{P_\rho}$ for some $0 < s \le 2$, we have

$$\mathbb{P}\left(\|u - u^{(m)}\|_V^2 \le \frac{C}{(m+m_0)^{s/(s+2)}}\right) \ge 1-\delta, \qquad 0 < \delta < 1, \quad m = 0,1,\ldots,$$

for some constant $C = C(M_\rho, \|u\|_{V^s_{P_\rho}}, m_0, s, \Lambda, \log(2/\delta)) < \infty$. Here, $V = H$ is an RKHS of functions $u : \Omega \to \mathbb{R}$ generated by some scalar-valued Mercer kernel $K : \Omega \times \Omega \to \mathbb{R}$ and $\Lambda = \max_{\omega\in\Omega} K(\omega,\omega)$. The associated maps R_ω are given by $R_\omega y = yK(\omega,\cdot)$, $y \in \mathbb{R}$. Consequently, $R^*_\omega u = u(\omega)$, $\omega \in \Omega$, corresponds to function evaluation. Thus, for $0 < s \le 1$, we get the same rate as in our Theorem 1 which, however, deals with the expectation of the squared error in $V = H$ in the more general vector-valued case. What our rather elementary method does not deliver is a result for the case $1 < s \le 2$ and for $L^2_\rho(\Omega,Y)$ convergence. For the latter situation, [**14**, Theorem C] gives the better estimate

$$\mathbb{P}\left(\|u-u^{(m)}\|^2_{L^2_\rho(\Omega,\mathbb{R})} \le \frac{\bar{C}}{(m+m_0)^{(s+1)/(s+2)}}\right) \ge 1-\delta, \quad 0<\delta<1, \quad m=0,1,\dots,$$

under the same assumptions but with a different constant

$$\bar{C} = \bar{C}(M_\rho, \|u\|_{V^s_{P_\rho}}, m_0, s, \Lambda, \log(2/\delta)) < \infty.$$

This is almost matching the lower estimates for kernel learning derived in [**1**] for classes of instances, where the spectrum of P_ρ exhibits a prescribed decay of the form $\lambda_k \asymp k^{-b}$ for some $b > 1$. Note that, for Mercer kernels, the operator

$$P_\rho :\ u \in H \quad \longmapsto \quad (P_\rho u)(\cdot) = \int_\Omega K(\cdot,\theta)u(\theta)\,d\rho(\theta)$$

is trace class whereas in our Theorem 1 no stronger decay of eigenvalues is assumed.

Estimates in expectation that are close to our result have also been obtained for slightly different settings. For example, in [**16**] both, the so-called *regularized* ($\alpha_m < 1$) and the *unregularized* online algorithm ($\alpha_m = 1$) were analyzed in the scalar-valued case $Y = \mathbb{R}$ under assumptions similar to ours for $L^2_\rho(\Omega,\mathbb{R})$ and $V = H$ convergence. We only quote the result for convergence in the RKHS $V = H$. It concerns the so-called *finite horizon* case of the unregularized online algorithm (14) with $\alpha_m = 1$, where one fixes $N < \infty$, chooses a constant step-size $\mu_m = \mu$, $m = 0,\dots,N-1$, which depends on N, stops the iteration at $m = N$ and asks for a good estimate of the expectation of $\mathbb{E}(\|u-u^{(N)}\|^2_V)$ for the final iterate only. Up to the specification of constants, Theorem 6 in [**16**] states that, under the condition $u \in V^s_{P_\rho}$, $s > 0$, one can achieve the bound

$$\mathbb{E}(\|u-u^{(N)}\|^2_V) = \mathrm{O}(N^{-s/(s+2)}), \qquad N \to \infty,$$

if one sets $\mu_N = cN^{-(s+1)/(s+2)}$ with a properly adjusted value of c. Note that $s > 0$ is arbitrary with the exponent approaching -1, if the smoothness parameter s tends to ∞, while our result does not provide improvements for $s > 1$. The drawback of the finite horizon case is that the estimate concerns only a fixed iterate $u^{(N)}$ with an N which needs to be decided on beforehand. In some sense, this can be viewed as building an approximation to the solution u_N of (12) with $\kappa_N = \mu_N$ from a single pass over the N i.i.d. samples (ω_m, y_m), $m = 0,\dots,N-1$.

In recent years, attention has shifted to obtaining refined rates when P_ρ possesses faster eigenvalue decay, usually expressed by the property that P^β_ρ is trace class for some $\beta < 1$ or by the slightly weaker assumption

$$\lambda_k = \mathrm{O}(k^{-1/\beta}), \qquad k \to \infty, \tag{17}$$

on the eigenvalues of the covariance operator P_ρ. Bounds involving knowledge about $\beta < 1$ are sometimes called capacity dependent, our bounds in Theorem 1 as well as the cited results from [**14**, **16**] are thus capacity independent. Capacity dependent convergence rates for the expected squared error for the online algorithm (14) have been obtained, among others, in [**6**,**7**,**9**,**10**], again in the scalar-valued case $Y = \mathbb{R}$ and with various parameter settings in (14), including unregularized and finite horizon versions. In [**7**], rates for $\mathbb{E}(\|u-\bar{u}^{(m)}\|^2_{L^2_\rho(\Omega,\mathbb{R})})$ have been established, where

$$\bar{u}^{(m)} = \frac{1}{m+1}\sum_{k=0}^{m} u^{(k)}, \qquad m = 0,1,\dots, \tag{18}$$

is the sequence of averages associated with the sequence $u^{(m)}$, $m = 0,1,\ldots$, obtained by the unregularized iteration (14) with $\alpha_m = 1$ and $u^{(0)} = 0$. That averaging has a similar effect as regularization with $\alpha_m = (m+1)/(m+2)$ in (14) considered in Theorem 1 can be guessed if one observes that

$$\bar{u}^{(m+1)} = \frac{m+1}{m+2}\bar{u}^{(m)} + \frac{1}{m+2}u^{(m+1)},$$

where $u^{(m+1)} = u^{(m)} + \mu_m(y_m - R_{\omega_m}u^{(m)})$, and compares with (14). To illustrate the influence of β, we formulate the following bound, which is a consequence of [**7**, Corollary 3]: Under an additional technical assumptions on the noise term ε_ω, if the condition (17) holds for some $0 < \beta < 1$ and $u \in V^s_{P_\rho}$, $s > -1$, then for suitable choices for the learning rates μ_m, we have

$$\mathbb{E}(\|u - \bar{u}^{(m)}\|^2_{L^2_\rho(\Omega,\mathbb{R})}) = \begin{cases} \mathrm{O}((m+1)^{-(s+1)}), & -1 < s < -\beta, \\ \mathrm{O}((m+1)^{-(s+1)/(s+1+\beta)}), & -\beta < s < 1-\beta, \\ \mathrm{O}((m+1)^{-(1-\beta/2)}), & 1-\beta < s. \end{cases}$$

Thus, stronger eigenvalue decay generally implies stronger asymptotic error decay in the $L^2_\rho(\Omega,\mathbb{R})$ norm. In [**6**, Section 6], similar rates are obtained in the finite horizon setting for both, the above averaged iterates $\bar{u}^{(N)}$ and for $u^{(N)}$ produced by a two-step extension of the one-step iteration (14).

In addition to $L^2_\rho(\Omega,\mathbb{R})$ convergence results, the paper [**10**] also provides a capacity dependent convergence estimate in the RKHS norm for the unregularized algorithm (14) with parameters $\alpha_m = 1$ and $\mu_m = c(m+1)^{-1/2}$. Under the boundedness assumption $|y| \le M_\rho$, Theorem 2 in [**10**] implies that

$$\mathbb{E}(\|u - u^{(m)}\|^2_V) = \mathrm{O}((m+1)^{-\min(s,1-\beta)/2)}\log^2(m+1)), \qquad m = 1,2,\ldots,$$

if $u \in V^s_{P_\rho}$ for some $s > 0$, P^β_ρ is trace class for some $0 < \beta < 1$, and c is properly adjusted.

Finally, the scalar-valued least-squares regression problem with $Y = \mathbb{R}$ and RKHS prior space H can also be cast as linear regression problem in $V = H$. This has been done in [**6**, **9**]. More abstractly, given a μ-distributed random variable $(\xi_\omega, y) \in V \times \mathbb{R}$ on $\Omega \times \mathbb{R}$, the task is to find approximations to the minimizer $u \in V$ of the problem

$$\mathbb{E}(|(\xi_\omega, v) - y|^2) \longmapsto \min, \qquad v \in V, \tag{19}$$

from i.i.d. samples (ξ_{ω_i}, y_i). If $V = H$ is the RKHS, which regularizes the scalar-valued least-squares regression problem on $\Omega \times \mathbb{R}$, then the canonical choice is $\xi_\omega = K(\omega,\cdot)$. In [**9**], for the iteration

$$u^{(m+1)} = u^{(m)} + \mu_m(y_m - (\xi_{\omega_m}, u^{(m)}))\xi_{\omega_m}, \qquad m = 0,1,\ldots,$$

weak convergence in V is studied by deriving estimates for quantities such as $\mathbb{E}((v, e^{(m)})^2)$ and $\mathbb{E}((\xi_{\omega'}, e^{(m)})^2)$ under some simplifying assumptions on the noise and the normalization $\|\xi_\omega\| = 1$. Note that this iteration is nothing but the unregularized iteration (14) with $\alpha_m = 1$ since $(\xi_{\omega_m}, u^{(m)}) = u^{(m)}(\omega_m)$ in this case. In the learning application, the assumption $\|\xi_\omega\| = 1$ means $K(\omega,\omega) = 1$. Moreover, in this case

$$\mathbb{E}((\xi_{\omega'}, e^{(m)})^2) = \mathbb{E}(\|u - u^{(m)}\|_{L^2_\rho(\Omega,\mathbb{R})}),$$

since the expectation on the left is, in addition to the i.i.d. samples (ξ_{ω_k}, y_k), $k = 0,\ldots,m-1$, also taken with respect to the independently ρ-distributed random variable $\xi_{\omega'}$. This links to learning rates in the $L^2_\rho(\Omega,\mathbb{R})$ norm. The estimates for $\mathbb{E}((\xi_{\omega'}, e^{(m)})^2)$ given in [**9**] concern both, the finite horizon and the online setting and again depend on the parameters $s \ge 0$ (smoothness of u) and $0 < \beta \le 1$ (capacity assumption on P_ρ). For the estimates of $\mathbb{E}((v, e^{(m)})^2)$, the smoothness $s' \ge 0$ of the fixed element $v \in V^{s'}_{P_\rho}$ is traded against the smoothness $s \ge 0$ of $u \in V^s_{P_\rho}$. We refer to [**9**] for the details.

4 Proof of Theorem 1

In this subsection, we will use the notation and assumptions outlined above, with the only change that the scalar product in V is simply denoted by $(\cdot,\cdot)$ and the associated norm $\|\cdot\|_V$ is accordingly denoted by $\|\cdot\|$. Moreover, recall that we have set $e^{(m)} = u - u^{(m)}$. We will prove an estimate of the form

$$\mathbb{E}(\|e^{(m)}\|^2) = \mathrm{O}((m+1)^{-s/(2+s)}), \qquad m \to \infty, \tag{20}$$

under the assumption $u \in V^s_{P_\rho}$, $0 < s \le 1$, if the parameters A and t in (15) are chosen accordingly. The precise statement and the dependence of the constant in (20) on initial error, noise variance and smoothness assumption are stated in the formulation of Theorem 1.

From (14) and $y_m = R^*_{\omega_m} u + \varepsilon_{\omega_m}$ we deduce the error representation

$$e^{(m+1)} = \underbrace{\alpha_m(e^{(m)} - \mu_m P_{\omega_m} e^{(m)}) + \bar{\alpha}_m u}_{\bar{e}^{(m+1)}:=} - \alpha_m \mu_m R_{\omega_m} \varepsilon_{\omega_m},$$

where $\bar{\alpha}_m := 1 - \alpha_m = (m+2)^{-1}$, compare also 15. The first term $\bar{e}^{(m+1)}$ corresponds to the noiseless case considered in [**8**] while the remainder term is the noise contribution. Thus,

$$\|e^{(m+1)}\|^2 = \|\bar{e}^{(m+1)}\|^2 - 2\alpha_m\mu_m(R_{\omega_m}\varepsilon_{\omega_m}, \bar{e}^{(m+1)}) + \alpha_m^2\mu_m^2\|R_{\omega_m}\varepsilon_{\omega_m}\|^2. \tag{21}$$

We now estimate the conditional expectation with respect to given $u^{(m)}$, separately for the three terms in (21). Here and in the following we denote this conditional expectation by $\mathbb{E}'$. For the third term, by (6) and the definition of the variance σ^2, we have

$$\mathbb{E}'(\|R_{\omega_m}\varepsilon_{\omega_m}\|^2) \le \Lambda \mathbb{E}(\|\varepsilon_\omega\|_Y^2) = \Lambda\sigma^2. \tag{22}$$

For the second term, we need

$$\mathbb{E}((R_\omega \varepsilon_\omega, w)) = \mathbb{E}((y - R^*_\omega u, R^*_\omega w)_Y) = 0 \qquad \forall\, w \in V.$$

This straightforwardly follows from the fact that $u \in V$ is the minimizer of the problem (9). Thus, by setting $w = \alpha_m e^{(m)} + \bar{\alpha}_m u$, we obtain

$$\begin{aligned}
&\mathbb{E}'(-2\alpha_m\mu_m(R_{\omega_m}\varepsilon_{\omega_m}, \bar{e}^{(m+1)}))\\
&\qquad = 2\alpha_m\mu_m(\alpha_m\mu_m\mathbb{E}'((R_{\omega_m}\varepsilon_{\omega_m}, P_{\omega_m}e^{(m)})) - \mathbb{E}'((R_{\omega_m}\varepsilon_{\omega_m}, w)))\\
&\qquad = 2\alpha_m^2\mu_m^2\mathbb{E}'((R_{\omega_m}\varepsilon_{\omega_m}, P_{\omega_m}e^{(m)}))\\
&\qquad \le \alpha_m^2\mu_m^2(\mathbb{E}'(\|R_{\omega_m}\varepsilon_{\omega_m}\|^2) + \mathbb{E}'(\|P_{\omega_m}e^{(m)}\|^2)).
\end{aligned}$$

Here, the first term is estimated by (22). For the second term, we substitute the upper bound

$$\mathbb{E}'(\|P_{\omega_m}e^{(m)}\|^2) \le \Lambda\mathbb{E}'((P_{\omega_m}e^{(m)}, e^{(m)})) = \Lambda(P_\rho e^{(m)}, e^{(m)}), \tag{23}$$

which follows from (6) and the definition of P_ρ. Together this gives

$$\mathbb{E}'(-2\alpha_m\mu_m(R_{\omega_m}\varepsilon_{\omega_m}, \bar{e}^{(m+1)})) \le \Lambda\alpha_m^2\mu_m^2(\sigma^2 + (P_\rho e^{(m)}, e^{(m)})) \tag{24}$$

for the second term in (21).

For the estimation of the first term $\mathbb{E}'(\|\bar{e}^{(m+1)}\|^2)$, we modify the arguments from [**8**], where the case $\varepsilon_m = 0$ was treated. We use the error decomposition

$$\begin{aligned}
\|\bar{e}^{(m+1)}\|^2 = \bar{\alpha}_m^2\|u\|^2 &+ 2\alpha_m\bar{\alpha}_m(u, e^{(m)} - \mu_m P_{\omega_m}e^{(m)}\\
&+ \alpha_m^2(\|e^{(m)}\|^2 - 2\mu_m(e^{(m)}, P_{\omega_m}e^{(m)})) + \mu_m^2\|P_{\omega_m}e^{(m)}\|^2).
\end{aligned}$$

After taking conditional expectations, we arrive with the definition of P_ρ and (23) at

$$\begin{aligned}\mathbb{E}'(\|\bar{e}^{(m+1)}\|^2) &= \bar{\alpha}_m^2\|u\|^2 + 2\alpha_m\bar{\alpha}_m(u, e^{(m)} - \mu_m P_\rho e^{(m)}) \\ &\quad + \alpha_m^2(\|e^{(m)}\|^2 - 2\mu_m(e^{(m)}, P_\rho e^{(m)}) + \mu_m^2\mathbb{E}'(\|P_{\omega_m}e^{(m)}\|^2)) \\ &\le \bar{\alpha}_m^2\|u\|^2 + 2\alpha_m\bar{\alpha}_m(u, e^{(m)} - \mu_m P_\rho e^{(m)}) \\ &\quad + \alpha_m^2(\|e^{(m)}\|^2 - \mu_m(2-\Lambda\mu_m)(e^{(m)}, P_\rho e^{(m)})).\end{aligned}$$

Next, in order to estimate the term $(u, e^{(m)} - \mu_m P_\rho e^{(m)})$, we take an arbitrary $h = P_\rho^{1/2} v \in V^1_{P_\rho}$, where $v \in V = V^0_{P_\rho}$ and $\|h\|_{V^1_{P_\rho}} = \|v\|$. With this, we have

$$\begin{aligned}&2\alpha_m\bar{\alpha}_m(u, e^{(m)} - \mu_m P_\rho e^{(m)}) \\ &\quad = 2\alpha_m\bar{\alpha}_m((u-h, (I-\mu_m P_\rho)e^{(m)}) + (h, (I-\mu_m P_\rho)e^{(m)})) \\ &\quad \le 2\alpha_m\bar{\alpha}_m\|u-h\|\|(I-\mu_m P_\rho)e^{(m)}\| + 2(\bar{\alpha}_m\mu_m^{-1/2}(I-\mu_m P_\rho)v, \alpha_m\mu_m^{1/2}e^{(m)}) \\ &\quad \le 2\alpha_m\bar{\alpha}_m\|u-h\|\|e^{(m)}\| + \bar{\alpha}_m^2\mu_m^{-1}\|(I-\mu_m P_\rho)v\|^2 + \alpha_m^2\mu_m\|P_\rho^{1/2}e^{(m)}\|^2 \\ &\quad \le 2\alpha_m\bar{\alpha}_m\|u-h\|\|e^{(m)}\| + \bar{\alpha}_m^2\mu_m^{-1}\|h\|^2_{V^1_{P_\rho}} + \alpha_m^2\mu_m(P_\rho e^{(m)}, e^{(m)}).\end{aligned}$$

Here, we have silently used that $\|(I-\mu_m P_\rho)e^{(m)}\| \le \|e^{(m)}\|$ and similarly

$$\|(I-\mu_m P_\rho)v\| \le \|v\| = \|h\|_{V^1_{P_\rho}},$$

which holds since $0 < \mu_m \le A \le (2\Lambda)^{-1}$ according to (15) and the restriction on A. Substitution into the previous inequality results in

$$\begin{aligned}\mathbb{E}'(\|\bar{e}^{(m+1)}\|^2) &\le \bar{\alpha}_m^2(\|u\|^2 + \mu_m^{-1}\|h\|^2_{V^1_{P_\rho}}) + 2\alpha_m\bar{\alpha}_m\|u-h\|\|e^{(m)}\| \\ &\quad + \alpha_m^2(\|e^{(m)}\|^2 - \mu_m(1-\Lambda\mu_m)(e^{(m)}, P_\rho e^{(m)})).\end{aligned}$$

Now, combining this estimate for the conditional expectation of the first term in (21) with the bounds (22) and (24) for the third and second term, respectively, we arrive at

$$\begin{aligned}\mathbb{E}'(\|\bar{e}^{(m+1)}\|^2) &\le \alpha_m^2(\|e^{(m)}\|^2 + 2\Lambda\sigma^2\mu_m^2) \\ &\quad + 2\alpha_m\bar{\alpha}_m\|u-h\|\|e^{(m)}\| + \bar{\alpha}_m^2(\|u\|^2 + \mu_m^{-1}\|h\|^2_{V^1_{P_\rho}}).\end{aligned} \tag{25}$$

Here, the term involving $(e^{(m)}, P_\rho e^{(m)}) \ge 0$ has been dropped since its forefactor $-\mu_m(1 - 2\Lambda\mu_m)$ is non-positive due to the restriction on A in (15).

For given

$$u = \sum_k c_k\psi_k \in V^s_{P_\rho}, \qquad 0 < s \le 1,$$

in (25) we choose

$$h = \sum_{k:\,\lambda_k(m+1)^b \ge B} c_k\psi_k$$

with some fixed constants $b, B > 0$ specified below. This gives

$$\|h\|^2_{V^1_{P_\rho}} = \sum_{k:\,\lambda_k(m+1)^b \ge B} \lambda_k^{-(1-s)}(\lambda_k^{-s}c_k^2) \le B^{-(1-s)}(m+1)^{(1-s)b}\|u\|^2_{V^s_{P_\rho}}$$

and

$$\|u-h\|^2 = \sum_{k:\,\lambda_k(m+1)^b < B} \lambda_k^s(\lambda_k^{-s}c_k^2) \le B^s(m+1)^{-bs}\|u\|^2_{V^s_{P_\rho}}.$$

After substitution into (25), we obtain

$$\mathbb{E}'(\|\bar{e}^{(m+1)}\|^2) \leq \alpha_m^2(\|e^{(m)}\|^2 + 2\Lambda\sigma^2\mu_m^2) + 2\alpha_m\bar{\alpha}_m B^{s/2}(m+1)^{-bs/2}\|u\|_{V^s_{P_\rho}}\|e^{(m)}\| \qquad (26)$$
$$+ \bar{\alpha}_m^2(\|u\|^2 + \mu_m^{-1}B^{-(1-s)}(m+1)^{(1-s)b}\|u\|^2_{V^s_{P_\rho}}).$$

Clearly, if $s=1$, we can set $h=u$ which would greatly simplify the considerations below and leads to a more precise final estimate, see Section 5.1.

Next, we switch to full expectations in (26) by using the independence assumption for the sampling process and take into account that

$$\varepsilon_m := \mathbb{E}(\|e^{(m)}\|^2)^{1/2} \geq \mathbb{E}(\|e^{(m)}\|).$$

Together with (15) and $\alpha_m = (m+1)\bar{\alpha}_m$, this gives

$$\begin{aligned}
\varepsilon_{m+1}^2 &\leq \alpha_m^2(\varepsilon_m^2 + 2A^2\Lambda\sigma^2(m+1)^{-2t}) + 2\alpha_m\bar{\alpha}_m B^{s/2}(m+1)^{-bs/2}\|u\|_{V^s_{P_\rho}}\varepsilon_m \\
&\quad + \bar{\alpha}_m^2(\|u\|^2 + A^{-1}B^{-(1-s)}(m+1)^{(1-s)b+t}\|u\|^2_{V^s_{P_\rho}}) \\
&\leq \alpha_m^2\varepsilon_m^2 + \bar{\alpha}_m^2(2A^2\Lambda\sigma^2(m+1)^{2-2t} + 2B^{s/2}(m+1)^{-bs/2+1}\|u\|_{V^s_{P_\rho}}\varepsilon_m\|e^{(m)}\| \\
&\quad + \|u\|^2 + A^{-1}B^{-(1-s)}(m+1)^{(1-s)b+t}\|u\|^2_{V^s_{P_\rho}}).
\end{aligned}$$

In a final step, we assume for a moment that

$$\varepsilon_k \leq C(k+1)^{-r}, \qquad k = 0,\dots,m, \qquad (27)$$

holds for some constants $C, r > 0$. Next, we set

$$a := \max(2-2t, -bs/2+1-r, (1-s)b+t)$$

and

$$D := 2A^2\Lambda\sigma^2 + 2CB^{s/2}\|u\|_{V^s_{P_\rho}} + \|u\|^2 + A^{-1}B^{-(1-s)}\|u\|^2_{V^s_{P_\rho}}.$$

Since $1/2 < t < 1$ is assumed in (15), we have $a > 0$. Then, for $k = 0,1,\dots,m$, the estimate for ε_{k+1} simplifies to

$$\varepsilon_{k+1}^2 \leq \alpha_k^2\varepsilon_k^2 + D\bar{\alpha}_k^2(k+1)^a$$

or, since $\alpha_k^2\bar{\alpha}_{k-1}^2 = \bar{\alpha}_k^2$, to

$$d_{k+1} := \bar{\alpha}_k^{-2}\varepsilon_{k+1}^2 \leq \alpha_k^2\bar{\alpha}_k^{-2}\varepsilon_k^2 + D(k+1)^a = d_k + D(k+1)^a.$$

By recursion we obtain

$$d_{m+1} \leq d_0 + D\sum_{k=0}^{m}(k+1)^a = \varepsilon_0^2 + D\sum_{k=0}^{m}(k+1)^a$$

and eventually

$$\varepsilon_{m+1}^2 \leq (m+2)^{-2}(\|e^{(0)}\|^2 + D(m+2)^{a+1}) < (\|e^{(0)}\|^2 + D)(m+2)^{a-1},$$

since we have $a > 0$ and

$$\sum_{k=0}^{m}(k+1)^a \leq \int_1^{m+2} x^a\,dx < (m+2)^{a+1}.$$

Thus, (27) holds by induction for all m if we ensure that

$$1-a \geq 2r, \qquad \|e^{(0)}\|^2 + D \leq C^2. \qquad (28)$$

To finish the proof of Theorem 1, it remains to maximize r for given $0 < s \le 1$. To this end, it is intuitively clear to require

$$a = 1 - 2r = 2 - 2t = -bs/2 + 1 - r = (1-s)b + t.$$

This system of equations has the unique solution

$$t = \frac{1+s}{2+s}, \quad b = \frac{1}{2+s}, \quad 2r = \frac{s}{2+s}, \quad a = \frac{2}{2+s}.$$

Furthermore, the appropriate value for C in (27) must satisfy

$$C^2 \ge \|e^{(0)}\|^2 + \|u\|^2 + 2A^2\Lambda\sigma^2 + 2CB^{s/2}\|u\|_{V^s_{P_\rho}} + A^{-1}B^{-(1-s)}\|u\|^2_{V^s_{P_\rho}}.$$

With such choices for t and C, the condition (28) is guaranteed and (27) yields the desired bound

$$\varepsilon_m^2 \le C^2(m+1)^{s/(s+2)}, \qquad m = 1, 2, \ldots, N-1.$$

By choosing concrete values for $0 < A \le (2\Lambda)^{-1}$ and $B > 0$, the constant C^2 can be made more explicit. E.g., substituting the upper bound

$$2CB^{s/2}\|u\|_{V^s_{P_\rho}} \le \frac{C^2}{2} + 2B^s\|u\|^2_{V^s_{P_\rho}}$$

and rearranging term shows that

$$C^2 = 2\left(\|e^{(0)}\|^2 + \|u\|^2 + B^s(2 + (AB)^{-1})\|u\|^2_{V^s_{P_\rho}} + 2A^2\Lambda\sigma^2\right)$$

is suitable. In particular, setting for simplicity A to its maximal value $A = (2\Lambda)^{-1}$ and taking $B = \Lambda$ gives a more explicit dependence of C^2 on the assumptions on $\|e^{(0)}\|^2$, the variance σ^2 and the smoothness of u, namely

$$C^2 = 2\|e^{(0)}\|^2 + 2\|u\|^2 + 8\Lambda^s\|u\|_{V^s_{P_\rho}} + \sigma^2/\Lambda. \tag{29}$$

This is the constant shown in the formulation of Theorem 1. Clearly, varying A and B will change the trade-off between initial error, noise variance and smoothness assumptions in the convergence estimate (27). Note also that B is not part of the algorithm and can be adjusted to any value. This finishes the proof of Theorem 1.

5 Further remarks

5.1 Comments on Theorem 1

In the special case $s = 1$, the proof of Theorem 2 simplifies as follows: In (25) we can set $h = u$ and (26) consequently simplifies to

$$\mathbb{E}'(\|\bar{e}^{(m+1)}\|^2) \le \alpha_m^2(\|e^{(m)}\|^2 + 2\Lambda\sigma^2\mu_m^2) + \bar{\alpha}_m^2(\|u\|^2 + \mu_m^{-1}\|u\|^2_{V^1_{P_\rho}}). \tag{30}$$

Thus, with $\mu_m = A(m+1)^{-t}$ we directly obtain a recursion for

$$d_m := \bar{\alpha}_{m-1}^{-2}\varepsilon_m^2 = (m+1)^2\mathbb{E}(\|e^{(m)}\|^2)$$

in the form

$$d_{m+1} \le d_m + (2A^2\Lambda\sigma^2(m+1)^{2-2t} + \|u\|^2 + A^{-1}(m+1)^t\|u\|^2_{V^1_{P_\rho}}).$$

For $1/2 < t < 1$ we finally arrive at

$$\mathbb{E}(\|e^{(m)}\|^2) \le \frac{\|e^{(0)}\|^2}{(m+1)^2} + \frac{2A^2\Lambda\sigma^2}{(m+1)^{2t-1}} + \frac{\|u\|^2}{m+1} + \frac{A^{-1}\|u\|^2_{V^1_{P_\rho}}}{(m+1)^{1-t}}, \tag{31}$$

$m = 1,2,\ldots$. This estimate shows more clearly the guaranteed error decay with respect to the initial error $\|e^{(0)}\|^2$, the noise variance σ^2 and the norms $\|u\|^2$ and $\|u\|^2_{V^1_{P_\rho}}$ of the solution u, respectively, in dependence on t. The asymptotically dominant term is here of the form

$$\mathrm{O}((m+1)^{-\min(2t-1,1-t)})$$

and is minimized if $t = 2/3$. For this value of t and with $A = (2\Lambda)^{-1}$ one obtains

$$\mathbb{E}(\|e^{(m)}\|^2) \le \frac{\|e^{(0)}\|^2}{(m+1)^2} + \frac{\|u\|^2}{m+1} + \frac{\|u\|^2_{V^1_{P_\rho}} + \sigma^2}{2\Lambda(m+1)^{1/3}}. \tag{32}$$

Without further assumptions, one cannot expect a better error decay rate, see Section 3 and Subsection 5.3.

Another comment concerns the finite horizon setting which is often treated instead of a true online method. Here one fixes a finite N, chooses a constant learning rate $\mu_m = \mu$ for $m = 0,\ldots,N-1$ in dependence on N, and asks for a best possible bound for $\mathbb{E}(\|e^{(N)}\|^2)$ only. Our approach easily delivers results for this case as well. We demonstrate this only for $s = 1$. For fixed $\mu_m = \mu$, the error recursion for the quantities d_m takes now the form

$$d_{m+1} \le d_m + (2A^2\Lambda\sigma^2(m+1)^2\mu^2 + \|u\|^2 + \mu^{-1}\|u\|^2_{V^1_{P_\rho}}), \quad m = 0,\ldots,N-1,$$

and gives

$$\mathbb{E}(\|e^{(N)}\|^2) \le \frac{\|e^{(0)}\|^2}{(N+1)^2} + 2\Lambda\sigma^2\mu^2(N+1) + \frac{\|u\|^2 + \mu^{-1}\|u\|^2_{V^1_{P_\rho}}}{N+1}.$$

Setting $\mu = A(N+1)^{-2/3}$ results in a final estimate for the finite horizon case similar to (32) but only for $m = N$.

There are obvious drawbacks of the whole setting in which Theorem 1 is formulated. First of all, the assumptions are qualitative at most: Since μ, and thus ρ, is usually not at our disposal, we cannot verify the assumption $u \in V^s_{P_\rho}$, nor assess the value of σ^2. Moreover, even though in view of the obtained results the restriction to learning rates μ_m of the form (15) may not cause issues, the choice of optimal values for t and A is by no means obvious. A rule for the adaptive choice of μ_m, which does not require knowledge about values for s and the size of norms of u but leads to the same quantitaive error decay as guaranteed by Theorem 1, would be desirable.

5.2 Difficulties with convergence in $L^2_\rho(\Omega,Y)$

Our result for the vector-valued case concerned convergence in V which is isometric to the RKHS H generated by $\mathbf{R}$. What we did not succeed in is to extend our methods to establish better asymptotic convergence rates of $f_{u^{(m)}} \to f_u$ in the $L^2_\rho(\Omega,Y)$ norm. It is not hard to see that, under the assumption (11) about the existence of the minimizer u in (9), error estimates in the $L^2_\rho(\Omega,Y)$ norm require the investigation of $\mathbb{E}(\|P^{1/2}_\rho e^{(m)}\|^2) = \mathbb{E}((P_\rho e^{(m)}, e^{(m)}))$ instead of $\mathbb{E}(\|e^{(m)}\|^2)$. If, in analogy with (21), one examines the error decomposition

$$\|P^{1/2}_\rho e^{(m+1)}\|^2 \le \|P^{1/2}_\rho \bar{e}^{(m+1)}\|^2 - 2\alpha_m\mu_m(P_\rho R_{\omega_m}\varepsilon_{\omega_m}, \bar{e}^{(m+1)}) + \alpha^2_m\mu^2_m\|P^{1/2}_\rho R_{\omega_m}\varepsilon_{\omega_m}\|^2,$$

then difficulties mostly arise from the first term in the right-hand side. Indeed, we have

$$\|P_\rho^{1/2}\bar{e}^{(m+1)}\|^2 = \bar{\alpha}_m^2\|P_\rho^{1/2}u\|^2 + 2\alpha_m\bar{\alpha}_m(P_\rho u, e^{(m)} - \mu_m P_{\omega_m}e^{(m)}) \\ + \alpha_m^2(\|P_\rho^{1/2}e^{(m)}\|^2 - 2\mu_m(P_\rho e^{(m)}, P_{\omega_m}e^{(m)})) + \mu_m^2\|P_\rho^{1/2}P_{\omega_m}e^{(m)}\|^2).$$

After taking conditional expectations $\mathbb{E}'(\|P_\rho^{1/2}\bar{e}^{(m+1)}\|^2)$, we obtain a negative term

$$-2\alpha_m^2\mu_m\|P_\rho e^{(m)}\|^2$$

on the right-hand side which needs to compensate for positive contributions from terms such as

$$\mathbb{E}'(\|P_\rho^{1/2}P_{\omega_m}e^{(m)}\|^2).$$

Since, in general, P_ρ does not commute with the operators P_ω, this strategy does not work without additional assumptions.

5.3 A special case

Let us now consider the particular "learning" problem of recovering an unknown element $u \in V$ from noisy measurements of its coefficients with respect to a CONS $\Psi = \{\psi_i\}_{i\in\mathbb{N}}$ in V by the online method considered in this paper. To this end, we assume that we are given μ-distributed random samples (i_m, y_m), where $i_m \in \mathbb{N}$ and

$$y_m = (u, \psi_{i_m}) + \varepsilon_m, \qquad m = 0, 1, \ldots \tag{33}$$

are the noisy samples of the coefficients (u, ψ_i). Starting from $u^{(0)} = 0$, we now want to approximate u by the iterates $u^{(m)}$ obtained from the online algorithm

$$u^{(m+1)} = \alpha_m u^{(m)} + \alpha_m\mu_m(y_m - (u^{(m)}, \psi_{i_m}))\psi_{i_m}, \qquad m = 0, 1, \ldots, \tag{34}$$

where the coefficients α_m and μ_m are given by (15) with $\Lambda = 1$. This is a special instance of (14) if we set $\Omega = \mathbb{N}$, $Y = \mathbb{R}$ and define $R_i : \mathbb{R} \to V$ and $R_i^* : V \to \mathbb{R}$ by $R_i y = y\psi_i$ and $R_i^* v = (v, \psi_i)$, respectively. To simplify things further, let i_m be i.i.d. samples from $\mathbb{N}$ with respect to a discrete probability measure ρ on $\mathbb{N}$ and let ε_m be i.i.d. random noise with zero mean and finite variance $\sigma^2 < \infty$ which is independent of i_m. This means that the underlying measure μ on $\mathbb{N} \times \mathbb{R}$ is a product measure. The associated operator P_ρ is given by

$$P_\rho v = \sum_{i\in\mathbb{N}} \rho_i(v, \psi_i)\psi_i,$$

its eigenvalues $\lambda_i = \rho_i$ are given by ρ, and it is trace class (w.l.o.g., we assume $\rho_1 \geq \rho_2 \geq \ldots$). The spaces $V_{P_\rho}^s$, $-\infty < s < \infty$, can now be identified as sets of formal orthogonal series

$$V_{P_\rho}^s := \left\{ u \sim \sum_{i\in\mathbb{N}} c_i\psi_i \ : \quad \|u\|_{V_{P_\rho}^s}^2 = \sum_{i\in\mathbb{N}} \rho_i^{-s}c_i^2 \right\}.$$

Obviously, $V_{P_\rho}^s \subset V = V_{P_\rho}^0$ for $s > 0$. Since functions $f : \mathbb{N} \to \mathbb{R}$ can be identified with formal series with respect to Ψ by

$$u \sim \sum_{i\in\mathbb{N}} c_i\psi_i \quad \leftrightarrow \quad f_u : f_u(i) = c_i,$$

we have $\|f_u\|_{L_\rho^2(\mathbb{N},\mathbb{R})} = \|u\|_{V_{P_\rho}^{-1}}$ and we can silently identify $L_\rho^2(\mathbb{N},\mathbb{R})$ with $V_{P_\rho}^{-1}$. Under the assumptions made, the underlying minimization problem (9) on V reads

$$\mathbb{E}(|f_v - y|^2) = \|f_v - f_u\|_{L_\rho^2(\mathbb{N},\mathbb{R})}^2 + \sigma^2 \quad \longmapsto \quad \min,$$

and, as expected, has u as its unique solution. This example also shows that it may sometimes be more appropriate to consider convergence in V than convergence in the sense of $L^2_\rho(\Omega, Y)$.

The simplicity of this example enables a rather comprehensive convergence theory with respect to the scale of $V^s_{P_\rho}$ spaces. We state the following results without detailed proof.

Theorem 2 *Let $-1 \le \bar{s} \le 0$, $s \ge 0$, and $\bar{s} < s \le \bar{s}+2$. Then, for the sampling process described above, the online algorithm (34) converges for $u \in V^s_{P_\rho}$ in the $V^{\bar{s}}_{P_\rho}$ norm with the bound*

$$\mathbb{E}(\|e^{(m)}\|^2_{V^{\bar{s}}_{P_\rho}}) \le C(m+1)^{-\min(\frac{s-\bar{s}}{s+2},\frac{2}{\bar{s}+4})}(A^{\bar{s}-s}\|u\|^2_{V^s_{P_\rho}} + A^{2+\bar{s}}\sigma^2), \quad m = 1,2,\ldots, \tag{35}$$

if the parameters t and A in (15) satisfy

$$t = t_{s,\bar{s}} := \max((s+1)/(s+2), (\bar{s}+3)/(\bar{s}+4)), \qquad 0 < A \le 1/2.$$

Setting $\bar{s} = 0$, one concludes from (35) that the convergence estimate for the online algorithm (14), which holds by Theorem 1 for $0 < s \le 1$ in the general case, is indeed matched. For $\bar{s} = -1$, which corresponds to $L^2_\rho(\mathbb{N},\mathbb{R})$ convergence, the rate is better and in line with known lower bounds.

The estimate (35) for the online algorithm (34) is best possible, in the sense that, under the conditions of Theorem 2, the exponent in (35) cannot be increased without additional assumptions on ρ. In particular, for $s > \bar{s}+2$ no further improvement is obtained, i.e., the estimate indeed saturates at $s = \bar{s}+2$. This can be seen from the following result.

Theorem 3 *Let $-1 \le \bar{s} \le 0 \le s < \infty$, $\bar{s} < s$ and $\sigma > 0$. For the online algorithm (34) we have*

$$\sup_{\rho} \sup_{u:\|u\|_{V^s_{P_\rho}}=1} (m+1)^{\min((s-\bar{s})/(2+s),2/(\bar{s}+4))}\mathbb{E}(\|e^{(m)}\|^2_{V^{\bar{s}}_{P_\rho}}) \ge c > 0, \tag{36}$$

$m = 1,2,\ldots,$, where c depends on $\bar{s}$, s, σ and the parameters t and A in (15), but is independent of m.

The proofs of these statements are elementary but rather tedious and will be given elsewhere. Let us just note that the simplicity of this example allows us to reduce the considerations to explicit linear recursions for expectations associated with the decomposition coefficients $c_i^{(m)} := (u^{(m)}, \psi_i)$ of the iterates $u^{(m)}$ with respect to Ψ for each $i \in \mathbb{N}$ separately. This is because

$$\mathbb{E}(\|e^{(m)}\|^2_{V^{\bar{s}}_{P_\rho}}) = \sum_i \rho_i^{-\bar{s}}\mathbb{E}((c_i^{(m)})^2), \qquad \|u\|^2_{V^s_{P_\rho}} = \|e^{(0)}\|^2_{V^s_{P_\rho}} = \sum_i \rho_i^{-s}c_i^2 \tag{37}$$

and

$$\begin{aligned} c_i^{(m+1)} &= \bar{\alpha}_m c_i + \alpha_m c_i^{(m)} + \alpha_m \begin{cases} \mu_m(y_{i_m} - (u^{(m)}, \psi_{i_m})), & i_m = i \\ 0, & i_m \ne i \end{cases} \\ &= \bar{\alpha}_m c_i + \alpha_m(c_i^{(m)} - \delta_{i,i_m}\mu_m(c_i^{(m)} + \varepsilon_m)) \end{aligned}$$

for $m = 0,1,\ldots$, where $\delta_{i,i_m} = 1$ with probability ρ_i, and $\delta_{i,i_m} = 0$ with probability $1-\rho_i$. Thus, if we denote $\varepsilon_{m,i} := \mathbb{E}((c_i^{(m)})^2)$ and $\bar{\varepsilon}_{m,i} := \mathbb{E}(c_i^{(m)})$ and use the independence assumption, we get a system of linear recursions

$$\begin{aligned} \varepsilon_{m+1,i} &= \alpha_m^2(1-\rho_i\mu_m(2-\mu_m))\varepsilon_{m,i} + 2\alpha_m\bar{\alpha}_m(1-\rho_i\mu_m)c_i\bar{\varepsilon}_{m,i} + \bar{\alpha}_m^2 c_i^2 + \rho_i\alpha_m^2\mu_m^2\sigma^2, \\ \bar{\varepsilon}_{m+1,i} &= \alpha_m(1-\rho_i\mu_m)\bar{\varepsilon}_{m,i} + \bar{\alpha}_m c_i, \end{aligned}$$

$m = 0,1,\ldots$, with starting values $\varepsilon_{0,i} = c_i^2$ and $\bar{\varepsilon}_{0,i} = c_i$. This system can, in principle, be solved explicitly. For instance, we straightforwardly have

$$\bar{\varepsilon}_{m,i} = \frac{1}{m+1}c_i S_m, \qquad S_m := \sum_{k=0}^{m} \Pi_k^{m-1},$$

where the notation

$$\Pi_k^{m-1} = (1-\frac{a}{m^t})\cdot\ldots(1-\frac{a}{(k+1)^t}),\quad 0\le k\le m-1,\quad \Pi_m^{m-1}=1, \tag{38}$$

is used with $a = A\rho_i$. Similarly, we get

$$\varepsilon_{m,i} \le \frac{1}{(m+1)^2}\left(2c_i^2\sum_{k=0}^{m}\Pi_k^{m-1}S_k + \rho_i\sigma^2\sum_{k=0}^{m-1}(k+1)^2\mu_k^2\Pi_{k+1}^{m-1}\right).$$

A matching lower bound for $\varepsilon_{m,i}$ can be obtained by using a slightly different value of a in the definition of the products Π_k^{m-1}. The remainder of the argument for Theorem 2 requires first the substitution of tight upper bounds for Π_k^{m-1} and S_k in dependence on t and a into the bounds for $\varepsilon_{m,i}$. Next, after substitution of the estimates for $\varepsilon_{m,i}$ into (37), the resulting series has to be further estimated separately for the index sets $I_1 := \{i : A\rho_i \le (m+1)^{t-1}\}$ and $I_2 := \mathbb{N}\backslash I_1$ followed by choosing the indicated optimal value of $t = t_{s,\bar{s}}$. This leads to the bound (35) in Theorem 2. For the proof of Theorem 3, lower bounds for Π_k^{m-1}, S_k and, consequently, for ε_m are needed, combined with choosing appropriate discrete probability distributions ρ.

Acknowledgement

Michael Griebel and Peter Oswald were supported by the *Hausdorff Center for Mathematics* in Bonn, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC-2047/1 - 390685813 and the CRC 1060 *The Mathematics of Emergent Effects* of the Deutsche Forschungsgemeinschaft.

References

1. Caponnetto, A., De Vito, A.: Optimal rates for the regularized least-squares algorithm. Found. Comput. Math. **7:3**, 331–368 (2007)
2. Carmeli, C., De Vito, E., Toigo, A.: Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. Anal. Appl. **4:4**, 377–408 (2006)
3. Carmeli, C., De Vito, E., Toigo, A., Umanita, V.: Vector valued reproducing kernel Hilbert spaces and universality. Anal. Appl. **8:1**, 19–61 (2010)
4. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bull. Amer. Math. Soc. **39:1**, 1–49 (2002)
5. Cucker, F., Zhou, D.-X.: Learning Theory: An Approximation Theory Viewpoint. Cambridge Univ. Press (2007)
6. Dieuleveut, A., Flammarion, N., Bach, F.: Harder, better, faster, stronger convergence rates for least-squares regression. J. Machine Learning Research **18**, 1–51 (2017)
7. Dieuleveut, A., Bach, F.: Nonparametric stochastic approximation with large step-sizes, Ann. Statistics **44:4**, 1363–1399 (2016)
8. Griebel, M., Oswald, P., Stochastic subspace correction in Hilbert space, Constr. Approx. **48:3**, 501–521 (2018)
9. Guo, X., Lin, J., Zhou, D.-X.: Rates of convergence of randomized Kaczmarz algorithms in Hilbert spaces. Appl. Comput. Harm. Anal. **61:3**, 288–318 (2022)
10. Guo, Z.-C., Shi, L.: Fast and strong convergence of online learning algorithms. Adv. Comput. Math. **45**, 2745–2770 (2019)
11. Micchelli, C., Pontil, M.: On learning vector-valued functions. Neural. Comput. **17:1**, 177–204 (2005)
12. Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press, Cambridge MA, (2002)
13. Steinwart, I., Christmann, A.: Support Vector Machines. Springer, New York (2008)
14. Tarrés, P., Yao, Y.: Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. IEEE Transactions on Information Theory **60:9**, 5716–5735 (2014)
15. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
16. Ying, Y., Pontil, M.: Online gradient descent learning algorithms. Found. Comput. Math. **8:5**, 561–596 (2008)