The Annals of Statistics
2016, Vol. 44, No. 4, 1363–1399
DOI: 10.1214/15-AOS1391

NONPARAMETRIC STOCHASTIC APPROXIMATION WITH LARGE STEP-SIZES[1]

BY AYMERIC DIEULEVEUT AND FRANCIS BACH

Département d'Informatique de l'Ecole Normale Supérieure

We consider the random-design least-squares regression problem within the reproducing kernel Hilbert space (RKHS) framework. Given a stream of independent and identically distributed input/output data, we aim to learn a regression function within an RKHS $\mathcal{H}$, even if the optimal predictor (i.e., the conditional expectation) is not in $\mathcal{H}$. In a stochastic approximation framework where the estimator is updated after each observation, we show that the averaged unregularized least-mean-square algorithm (a form of stochastic gradient descent), given a sufficient large step-size, attains optimal rates of convergence for a variety of regimes for the smoothnesses of the optimal prediction function and the functions in $\mathcal{H}$. Our results apply as well in the usual finite-dimensional setting of parametric least-squares regression, showing adaptivity of our estimator to the spectral decay of the covariance matrix of the covariates.

1. Introduction. Positive-definite-kernel-based methods such as the support vector machine or kernel ridge regression are now widely used in many areas of science of engineering. They were first developed within the statistics community for nonparametric regression using splines, Sobolev spaces, and more generally reproducing kernel Hilbert spaces (see, e.g., [43]). Within the machine learning community, they were extended in several interesting ways (see, e.g., [33, 35]): (a) other problems were tackled using positive-definite kernels beyond regression problem, through the "kernelization" of classical unsupervised learning methods such as principal component analysis, canonical correlation analysis, or K-means, (b) efficient algorithms based on convex optimization have emerged, in particular for large sample sizes and (c) kernels for nonvectorial data have been designed for objects like strings, graphs, measures, etc. A key feature is that they allow the separation of the representation problem (designing good kernels for nonvectorial data) and the algorithmic/theoretical problems (given a kernel, how to design, run efficiently and analyse estimation algorithms).

The theoretical analysis of nonparametric least-squares regression within the RKHS framework is well understood. In particular, regression on input data in $\mathbb{R}^d$, $d \geq 1$, and so-called *Mercer kernels* (continuous kernels over a compact set) that

Received September 2014; revised July 2015.
[1]Supported in part by the European Research Council (SIERRA Project).
MSC2010 subject classifications. 60K35.
Key words and phrases. Reproducing kernel Hilbert space, stochastic approximation.

lead to dense subspaces of the space of square-integrable functions and nonparametric estimation [41], has been widely studied in the last decade starting with the works of Smale and Cucker [11, 12] and being further refined [14, 36] up to optimal rates [3, 10, 38] for Tikhonov regularization (batch iterative methods were for their part studied in [7, 30]). However, the kernel framework goes beyond Mercer kernels and nonparametric regression; indeed, kernels on nonvectorial data provide examples where the usual topological assumptions may not be natural, such as sequences, graphs and measures. Moreover, even finite-dimensional Hilbert spaces may need a more refined analysis when the dimension of the Hilbert space is much larger than the number of observations: for example, in modern text and web applications, linear predictions are performed with a large number of covariates which are equal to zero with high probability. The sparsity of the representation allows to reduce significantly the complexity of traditional optimization procedures; however, the finite-dimensional analysis which ignores the spectral structure of the data often leads to trivial guarantees because the number of covariates far exceeds the number of observations, while the analysis we carry out is meaningful (note that in these contexts sparsity of the underlying estimator is typically not a relevant assumption). In this paper, we consider minimal assumptions regarding the input space and the distributions, so that our nonasymptotic results may be applied to all the cases mentioned above.

In practice, estimation algorithms based on regularized empirical risk minimization (e.g., penalized least-squares) face two challenges: (a) using the correct regularization parameter and (b) finding an approximate solution of the convex optimization problem. In this paper, we consider these two problems jointly by following a stochastic approximation framework formulated directly in the RKHS, in which each observation is used only once and overfitting is avoided by making only a single pass through the data—a form of *early stopping*, which has been considered in other statistical frameworks such as boosting [49]. While this framework has been considered before [32, 39, 47], the algorithms that are considered either (a) require two sequences of hyperparameters (the step-size in stochastic gradient descent and a regularization parameter) or (b) do not always attain the optimal rates of convergence for estimating the regression function. In this paper, we aim to remove simultaneously these two limitations.

Traditional online stochastic approximation algorithms, as introduced by Robbins and Monro [31], lead in finite-dimensional learning problems (e.g., parametric least-squares regression) to stochastic gradient descent methods with step-sizes decreasing with the number of observations n, which are typically proportional to $n^{-\zeta}$, with ζ between $1/2$ and 1. Short step-sizes ($\zeta = 1$) are adapted to well-conditioned problems (low dimension, low correlations between covariates), while longer step-sizes ($\zeta = 1/2$) are adapted to ill-conditioned problems (high dimension, high correlations) but with a worse convergence rate; see, for example, [4, 34] and references therein. More recently, [5] showed that constant step-sizes *with averaging* could lead to the best possible convergence rate in Euclidean spaces (i.e.,

in finite dimensions). In this paper, we show that using longer step-sizes with averaging also brings benefits to Hilbert space settings needed for nonparametric regression.

With our analysis, based on positive definite kernels, under assumptions on both the objective function and the covariance operator of the RKHS, we derive improved rates of convergence [10], in both the finite horizon setting where the number of observations is known in advance and our bounds hold for the last iterate (with exact constants), and the online setting where our bounds hold for each iterate (asymptotic results only). It leads to an explicit choice of the step-sizes (which play the role of the regularization parameters) which may be used in stochastic gradient descent, depending on the number of training examples we want to use and on the assumptions we make.

In this paper, we make the following contributions:

- We review in Section 2 a general though simple algebraic framework for least-squares regression in RKHS, which encompasses all commonly encountered situations. This framework, however, makes unnecessary topological assumptions, which we relax in Section 2.5 (with details in Appendix A).
- We characterize in Section 3 the convergence rate of averaged least-mean-squares (LMS) and show how the proper set-up of the step-size leads to optimal convergence rates (as they were proved in [10]), extending results from finite-dimensional [5] to infinite-dimensional settings. The problem we solve here was stated as an open problem in [32, 47]. Moreover, our results apply as well in the usual finite-dimensional setting of parametric least-squares regression, showing adaptivity of our estimator to the spectral decay of the covariance matrix of the covariates (see Section 4.1).
- We compare our new results with existing work, both in terms of rates of convergence in Section 4, and with simulations on synthetic spline smoothing in Section 5.

Sketches of the proofs are given in Appendix B.
Complete proofs are available in the arXiv version of the paper [15].

2. Learning with positive-definite kernels. In this paper, we consider a general random design regression problem, where observations (x_i, y_i) are independent and identically distributed (i.i.d.) random variables in $\mathcal{X} \times \mathcal{Y}$ drawn from a probability measure ρ on $\mathcal{X} \times \mathcal{Y}$. The set $\mathcal{X}$ may be any set equipped with a measure; moreover, we consider for simplicity $\mathcal{Y} = \mathbb{R}$ and we measure the risk of a function $g : \mathcal{X} \to \mathbb{R}$, by the mean square error, that is, $\varepsilon(g) := \mathbb{E}_\rho[(g(X) - Y)^2]$.

The function g that minimizes $\varepsilon(g)$ over all measurable functions is known to be the conditional expectation, that is, $g_\rho(X) = \mathbb{E}[Y|X]$. In this paper, we consider formulations where our estimates lie in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

2.1. *Reproducing kernel Hilbert spaces.* Throughout this section, we make the following assumption:

(A1) $\mathcal{X}$ is a compact topological space and $\mathcal{H}$ is an RKHS associated with a continuous kernel K on the set $\mathcal{X}$.

RKHSs are well-studied Hilbert spaces which are particularly adapted to regression problems (see, e.g., [6, 43]). They satisfy the following properties:

1. $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is a separable Hilbert space of functions: $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$.
2. $\mathcal{H}$ contains all functions $K_x : t \mapsto K(x,t)$, for all x in $\mathcal{X}$.
3. For any $x \in \mathcal{X}$ and $f \in \mathcal{H}$, the reproducing property holds:

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}}.$$

The reproducing property allows to treat nonparametric estimation in the same algebraic framework as parametric regression. The Hilbert space $\mathcal{H}$ is totally characterized by the positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which simply needs to be a symmetric function on $\mathcal{X} \times \mathcal{X}$ such that for any finite family of points $(x_i)_{i \in I}$ in $\mathcal{X}$, the $|I| \times |I|$-matrix of kernel evaluations is positive semi-definite. We provide examples in Section 2.6. For simplicity, we have here made the assumption that K is a Mercer kernel, that is, $\mathcal{X}$ is a compact set and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous. See Section 2.5 for an extension without topological assumptions.

2.2. *Random variables.* In this paper, we consider a set $\mathcal{X}$ and $\mathcal{Y} \subset \mathbb{R}$ and a distribution ρ on $\mathcal{X} \times \mathcal{Y}$. We denote by ρ_X the marginal law on the space $\mathcal{X}$ and by $\rho_{Y|X=x}$ the conditional probability measure on Y given $x \in \mathcal{X}$. We may use the notation $\mathbb{E}[f(X)]$ or $\mathbb{E}_{\rho_X}[f(\cdot)]$ for $\int_{\mathcal{X}} f(x)\, d\rho_X(x)$. Beyond the moment conditions stated below, we will always make the assumptions that the space $L^2_{\rho_X}$ of square ρ_X-integrable functions defined below is separable (this is the case in most interesting situations; see [40] for more details). Since we will assume that ρ_X has full support, we will make the usual simplifying identification of functions and their equivalence classes (based on equality up to a zero-measure set). We denote by $\| \cdot \|_{L^2_{\rho_X}}$ the norm

$$\|f\|^2_{L^2_{\rho_X}} = \int_{\mathcal{X}} |f(x)|^2 \, d\rho_X(x).$$

The space $L^2_{\rho_X}$ is then a Hilbert space with norm $\| \cdot \|_{L^2_{\rho_X}}$, which we will always assume separable (i.e., with a countable orthonormal system).

Throughout this section, we make the following simple assumption regarding finiteness of moments:

(A2) $R^2 := \sup_{x \in \mathcal{X}} K(x,x)$ and $\mathbb{E}[Y^2]$ are finite; ρ_X has full support in $\mathcal{X}$.

Note that under these assumptions, any function in $\mathcal{H}$ is in $L^2_{\rho_X}$; however, this inclusion is strict in most interesting situations.

2.3. *Minimization problem.* We are interested in minimizing the following quantity, which is the *prediction error* (or mean squared error) of a function f, defined for any function in $L^2_{\rho_X}$ as

$$\varepsilon(f) = \mathbb{E}\big[\big(f(X) - Y\big)^2\big]. \tag{2.1}$$

We are looking for a function with a low prediction error in the particular function space $\mathcal{H}$, that is, we aim to minimize $\varepsilon(f)$ over $f \in \mathcal{H}$. We have for $f \in L^2_{\rho_X}$,

$$\begin{aligned}\varepsilon(f) &= \|f\|^2_{L^2_{\rho_X}} - 2\left\langle f, \int_{\mathcal{Y}} y \, d\rho_{Y|X=\cdot}(y)\right\rangle_{L^2_{\rho_X}} + \mathbb{E}\big[Y^2\big] \\ &= \|f\|^2_{L^2_{\rho_X}} - 2\big\langle f, \mathbb{E}[Y|X=\cdot]\big\rangle_{L^2_{\rho_X}} + \mathbb{E}\big[Y^2\big].\end{aligned} \tag{2.2}$$

A minimizer g of $\varepsilon(g)$ over $L^2_{\rho_X}$ is known to be such that $g(X) = \mathbb{E}[Y|X]$. Such a function is generally referred to as the regression function, and denoted g_ρ as it only depends on ρ. It is moreover unique (as an element of $L^2_{\rho_X}$). An important property of the prediction error is that the excess risk may be expressed as a squared distance to g_ρ, that is,

$$\forall f \in L^2_{\rho_X}, \qquad \varepsilon(f) - \varepsilon(g_\rho) = \|f - g_\rho\|^2_{L^2_{\rho_X}}. \tag{2.3}$$

A key feature of our analysis is that we only considered $\|f - g_\rho\|^2_{L^2_{\rho_X}}$ as a measure of performance and do not consider convergences in stricter norms (which are not true in general). *This allows us to neither assume that g_ρ is in $\mathcal{H}$ nor that $\mathcal{H}$ is dense in $L^2_{\rho_X}$.* We thus need to define a notion of the best estimator in $\mathcal{H}$. We first define the closure $\overline{F}$ (with respect to $\|\cdot\|_{L^2_{\rho_X}}$) of any set $F \subset L^2_{\rho_X}$ as the set of limits in $L^2_{\rho_X}$ of sequences in F. The space $\overline{\mathcal{H}}$ is a closed and convex subset in $L^2_{\rho_X}$. We can thus define $g_{\mathcal{H}} = \arg\min_{f \in \overline{\mathcal{H}}} \varepsilon(g)$, as the orthogonal projection of g_ρ on $\overline{\mathcal{H}}$, using the existence of the projection on any closed convex set in a Hilbert space. See Proposition 8 in Appendix A for details. Of course, we do not have $g_{\mathcal{H}} \in \mathcal{H}$, that is *the minimum in $\mathcal{H}$ is in general not attained.*

Estimation from n i.i.d. observations builds a sequence $(g_n)_{n\in\mathbb{N}}$ in $\mathcal{H}$. We will prove under suitable conditions that such an estimator satisfies weak consistency, that is, g_n ends up predicting as well as $g_{\mathcal{H}}$:

$$\mathbb{E}\big[\varepsilon(g_n) - \varepsilon(g_{\mathcal{H}})\big] \overset{n\to\infty}{\longrightarrow} 0 \quad \Leftrightarrow \quad \|g_n - g_{\mathcal{H}}\|_{\mathcal{L}^2_{\rho_X}} \overset{n\to\infty}{\longrightarrow} 0.$$

Seen as a function of $f \in \mathcal{H}$, our loss function ε is not coercive (i.e., not strongly convex), as our covariance operator (see definition below) Σ has no minimal strictly positive eigenvalue (the sequence of eigenvalues decreases to zero). As a consequence, even if $g_{\mathcal{H}} \in \mathcal{H}$, g_n may not converge to $g_{\mathcal{H}}$ in $\mathcal{H}$, and *when $g_{\mathcal{H}} \notin \mathcal{H}$, we shall even have $\|g_n\|_{\mathcal{H}} \to \infty$.*

2.4. *Covariance operator.* We now define the *covariance operator* for the space $\mathcal{H}$ and probability distribution ρ_X. The spectral properties of such an operator have appeared to be a key point to characterize the convergence rates of estimators [10, 11, 36].

We implicitly define (via Riesz' representation theorem) a linear operator $\Sigma : \mathcal{H} \to \mathcal{H}$ through

$$\forall (f, g) \in \mathcal{H}^2, \qquad \langle f, \Sigma g \rangle_{\mathcal{H}} = \mathbb{E}\big[f(X)g(X)\big] = \int_{\mathcal{X}} f(x)g(x)\, d\rho_X(x).$$

This operator is the *covariance operator* (defined on the Hilbert space $\mathcal{H}$). Using the reproducing property, we have

$$\Sigma = \mathbb{E}[K_X \otimes K_X],$$

where for any elements $g, h \in \mathcal{H}$, we denote by $g \otimes h$ the operator from $\mathcal{H}$ to $\mathcal{H}$ defined as

$$g \otimes h \ : f \mapsto \langle f, h \rangle_{\mathcal{H}} g.$$

Note that this expectation is formally defined as a Bochner expectation (an extension of Lebesgue integration theory to Banach spaces, see [28]) in $\mathcal{L}(\mathcal{H})$ the set of endomorphisms of $\mathcal{H}$.

In finite dimension, that is, $\mathcal{H} = \mathbb{R}^d$, for $g, h \in \mathbb{R}^d$, $g \otimes h$ may be identified to a rank-one matrix, that is, $g \otimes h = gh^\top = ((g_i h_j)_{1 \le i, j \le d}) \in \mathbb{R}^{d \times d}$ as for any f, $(gh^\top) f = g(h^\top f) = \langle f, h \rangle_{\mathcal{H}} g$. In other words, $g \otimes h$ is a linear operator, whose image is included in $\mathrm{Vect}(g)$, the linear space spanned by g. Thus in finite dimension, Σ is the usual (noncentered) covariance matrix.

We have defined the covariance operator on the Hilbert space $\mathcal{H}$. If $f \in \mathcal{H}$, we have for all $z \in \mathcal{X}$, using the reproducing property

$$\mathbb{E}\big[f(X)K(X, z)\big] = \mathbb{E}\big[f(X)K_z(X)\big] = \langle K_z, \Sigma f \rangle_{\mathcal{H}} = (\Sigma f)(z),$$

which shows that the operator Σ may be extended to any square-integrable function $f \in L^2_{\rho_X}$. In the following, we extend such an operator as an endomorphism T from $L^2_{\rho_X}$ to $L^2_{\rho_X}$.

DEFINITION 1 (Extended covariance operator). Assume (A1)–(A2). We define the operator T as follows:

$$T : L^2_{\rho_X} \to L^2_{\rho_X},$$
$$g \mapsto \int_{\mathcal{X}} g(t) K_t \, d\rho_{\mathcal{X}}(t),$$

so that for any $z \in \mathcal{X}$, $T(g)(z) = \int_{\mathcal{X}} g(x)K(x, z)\, d\rho_{\mathcal{X}}(t) = \mathbb{E}[g(X)K(X, z)]$.

From the discussion above, if $f \in \mathcal{H} \subset L^2_{\rho_X}$, then $Tf = \Sigma f$. We give here some of the most important properties of T. The operator T (which is an endomorphism of the separable Hilbert space $L^2_{\rho_X}$) may be reduced in some Hilbertian eigenbasis of $L^2_{\rho_X}$. It allows us to define the power of such an operator T^r, which will be used to quantify the regularity of the function $g_{\mathcal{H}}$. See the proof in Appendix I.2, Proposition 19 in [15].

PROPOSITION 1 (Eigendecomposition of T). *Assume* (A1)–(A2). *T is a bounded self-adjoint semi-definite positive operator on $L^2_{\rho_X}$, which is trace-class. There exists a Hilbertian eigenbasis $(\phi_i)_{i\in I}$ of the orthogonal supplement S of the null space* $\operatorname{Ker}(T)$, *with summable strictly positive eigenvalues $(\mu_i)_{i\in I}$. That is*:

- $\forall i \in I, T\phi_i = \mu_i\phi_i$, $(\mu_i)_{i\in I}$ *strictly positive such that* $\sum_{i\in I} \mu_i < \infty$.
- $L^2_{\rho_X} = \operatorname{Ker}(T) \overset{\perp}{\oplus} S$, *that is*, $L^2_{\rho_X}$ *is the orthogonal direct sum of* $\operatorname{Ker}(T)$ *and* S.

When the space S has finite dimension, then I has finite cardinality, while in general I is countable. Moreover, the null space $\operatorname{Ker}(T)$ may be either reduced to $\{0\}$ (this is the more classical setting and such an assumption is often made), finite-dimensional (e.g., when the kernel has zero mean, thus constant functions are in S) or infinite-dimensional (e.g., when the kernel space only consists in even functions, the whole space of odd functions is in S).

Moreover, the linear operator T allows to relate $L^2_{\rho_X}$ and $\mathcal{H}$ in a very precise way. For example, when $g \in \mathcal{H}$, we immediately have $Tg = \Sigma g \in \mathcal{H}$ and $\langle g, Tg\rangle_{\mathcal{H}} = \mathbb{E}g(X)^2 = \|g\|^2_{L^2_{\rho_X}}$. As we formally state in the following propositions, this essentially means that $T^{1/2}$ will be an isometry from $L^2_{\rho_X}$ to $\mathcal{H}$. We first show that the linear operator T happens to have an image included in $\mathcal{H}$, and that the eigenbasis of T in $L^2_{\rho_X}$ may also be seen as the eigenbasis of Σ in $\mathcal{H}$ (see the proof in Appendix I.2, Proposition 19 in [15]).

PROPOSITION 2 (Decomposition of Σ). *Assume* (A1)–(A2). *$\Sigma : \mathcal{H} \to \mathcal{H}$ is injective. The image of T is included in $\mathcal{H}$*: $\operatorname{Im}(T) \subset \mathcal{H}$. *Moreover, for any $i \in I$, $\phi_i = \frac{1}{\mu_i} T\phi_i \in \mathcal{H}$, thus $(\mu_i^{1/2}\phi_i)_{i\in I}$ is an orthonormal eigensystem of Σ and an Hilbertian basis of $\mathcal{H}$, that is, for any i in I, $\Sigma\phi_i = \mu_i\phi_i$.*

This proposition will be generalized under relaxed assumptions (in particular as Σ will no more be injective; see Section 2.5 and Appendix A). It means that the orthonormal system covers $\mathcal{H}$, as any function in $\mathcal{H}$ which would not be in $\operatorname{span}\{\Phi_i, i \in I\}$, would be in the null space of Σ.

We may now define all powers T^r (they are always well defined because the sequence of eigenvalues is upper-bounded).

DEFINITION 2 (Powers of T). We define, for any $r \geq 0$, $T^r : L^2_{\rho_X} \to L^2_{\rho_X}$, for any $h \in \operatorname{Ker}(T)$ and $(a_i)_{i\in I}$ such that $\sum_{i\in I} a_i^2 < \infty$, through: $T^r(h + \sum_{i\in I} a_i\phi_i) = \sum_{i\in I} a_i \mu_i^r \phi_i$. Moreover, for any $r > 0$, T^r may be defined as a bijection from S into $\operatorname{Im}(T^r)$. We may thus define its unique inverse $T^{-r} : \operatorname{Im}(T^r) \to S$.

The following proposition is a consequence of Mercer's theorem [2, 11]. It describes how the space $\mathcal{H}$ is related to the image of operator $T^{1/2}$.

PROPOSITION 3 (Isometry for Mercer kernels). *Under assumptions* (A1), (A2), $\mathcal{H} = T^{1/2}(L^2_{\rho_X})$ *and* $T^{1/2} : S \to \mathcal{H}$ *is an isometrical isomorphism.*

The proposition has the following consequences.

COROLLARY 1. *Assume* (A1), (A2):

- *For any* $r \geq 1/2$, $T^r(S) \subset \mathcal{H}$, *because* $T^r(S) \subset T^{1/2}(S)$, *that is, with large enough powers* r, *the image of* T^r *is in the Hilbert space.*
- $\forall r > 0$, $\overline{T^r(L^2_{\rho_X})} = S = \overline{T^{1/2}(L^2_{\rho_X})} = \overline{\mathcal{H}}$, *because* (a) $T^{1/2}(L^2_{\rho_X}) = \mathcal{H}$ *and* (b) *for any* $r > 0$, $\overline{T^r(L^2_{\rho_X})} = S$. *In other words, elements of* $\overline{\mathcal{H}}$ (*on which our minimization problem attains its minimum*), *may seen as limits* (*in* $L^2_{\rho_X}$) *of elements of* $T^r(L^2_{\rho_X})$, *for any* $r > 0$.
- $\mathcal{H}$ *is dense in* $L^2_{\rho_X}$ *if and only if* T *is injective* [*which is equivalent to* $\ker(T) = \{0\}$].

The sequence of spaces $\{T^r(L^2_{\rho_X})\}_{r>0}$ is thus a decreasing (when r is increasing) sequence of subspaces of $L^2_{\rho_X}$ such that any of them is dense in $\overline{\mathcal{H}}$, and $T^r(L^2_{\rho_X}) \subset \mathcal{H}$ if and only if $r \geq 1/2$.

In the following, the regularity of the function $g_{\mathcal{H}}$ will be characterized by the fact that $g_{\mathcal{H}}$ belongs to the space $T^r(L^2_{\rho_X})$ (and not only to its closure), for a specific $r > 0$ (see Section 2.7). This space may be described through the eigenvalues and eigenvectors as

$$T^r\big(L^2_{\rho_X}\big) = \left\{ \sum_{i=1}^{\infty} b_i \phi_i \text{ such that } \sum_{i=1}^{\infty} \frac{b_i^2}{\mu_i^{2r}} < \infty \right\}.$$

We may thus see the spaces $T^r(L^2_{\rho_X})$ as spaces of sequences with various decay conditions.

2.5. *Minimal assumptions.* In this section, we describe under which "minimal" assumptions the analysis may be carried. We prove that the set $\mathcal{X}$ may only be assumed to be equipped with a measure, the kernel K may only assumed to have bounded expectation $\mathbb{E}_\rho K(X, X)$ and the output Y may only be assumed to have finite variance. That is:

(A1′) $\mathcal{H}$ is a separable RKHS associated with kernel K on the set $\mathcal{X}$.
(A2′) $\mathbb{E}[K(X,X)]$ and $\mathbb{E}[Y^2]$ are finite.

In this section, we have to distinguish the set of square ρ_X-integrable functions $\mathcal{L}^2_{\rho_X}$ and its quotient $L^2_{\rho_X}$ that makes it a separable Hilbert space. We define p the projection from $\mathcal{L}^2_{\rho_X}$ into $L^2_{\rho_X}$ (precise definitions are given in Appendix A). Indeed it is no more possible to identify the space $\mathcal{H}$, which is a subset of $\mathcal{L}^2_{\rho_X}$, and its canonical projection $p(\mathcal{H})$ in $L^2_{\rho_X}$.

Minimality: The separability assumption is necessary to be able to expand any element as an infinite sum, using a countable orthonormal family (this assumption is satisfied in almost all cases, e.g., it is simple as soon as $\mathcal{X}$ admits a topology for which it is separable and functions in $\mathcal{H}$ are continuous; see [6] for more details). Note that we do not make any topological assumptions regarding the set $\mathcal{X}$. We only assume that it is equipped with a probability measure.

Assumption (A2′) is needed to ensure that every function in $\mathcal{H}$ is square-integrable, that is, $\mathbb{E}[K(X,X)] < \infty$ if and only if $\mathcal{H} \subset \mathcal{L}^2_{\rho_X}$; for example, for $f = K_z$, $z \in \mathcal{X}$, $\|K_z\|^2_{L^2_{\rho_X}} = \mathbb{E}[K(X,z)^2] \le K(z,z)\mathbb{E}K(X,X)$ (see more details in Appendix I, Proposition 11 in [15]).

Our assumptions are sufficient to analyze the minimization of $\varepsilon(f)$ with respect to $f \in \mathcal{H}$ and seem to allow the widest generality.

Comparison: These assumptions will include the previous setting, but also recover measures without full support (e.g., when the data live in a small subspace of the whole space) and kernels on discrete objects (with nonfinite cardinality).

Moreover, (A1′), (A2′) are strictly weaker than (A1), (A2). In previous work, (A2′) was sometimes replaced by the stronger assumptions $\sup_{x\in\mathcal{X}} K(x,x) < \infty$ [32, 39, 47] and $|Y|$ bounded [32, 39]. Note that in functional analysis, the weaker hypothesis $\int_{\mathcal{X}\times\mathcal{X}} k(x,x')^2 \, d\rho_X(x) \, d\rho_X(x') < \infty$ is often used [9], but it is not adapted to the statistical setting.

Main differences: The main difference here is that we cannot identify $\mathcal{H}$ and $p(\mathcal{H})$, that is, there may exist functions $f \in \mathcal{H} \setminus \{0\}$ such that $\|f\|_{\mathcal{L}^2_{\rho_X}} = 0$. This may, for example, occur if the support of ρ_X is strictly included in $\mathcal{X}$, and f is zero on this support, but not identically zero. See Appendix I.5 in [15] for more details.

As a consequence, Σ is no more injective and we do not have $\operatorname{Im}(T^{1/2}) = \mathcal{H}$ any more. We thus denote $\mathscr{S}$ an orthogonal supplement of the null space $\operatorname{Ker}(\Sigma)$. As we also need to be careful not to confuse $\mathcal{L}^2_{\rho_X}$ and $L^2_{\rho_X}$, we define an extension $\mathcal{T}$ of Σ from $\mathcal{L}^2_{\rho_X}$ into $\mathcal{H}$, then $T = p \circ \mathcal{T}$. We can define for $r \ge 1/2$ the power operator $\mathcal{T}^r$ of $\mathcal{T}$ (from $L^2_{\rho_X}$ into $\mathcal{H}$); see Appendix A for details.

Conclusion: Our problem has the same behavior under such assumptions. Proposition 1 remains unchanged. Decompositions in Proposition 2 and Corollary 1 must be slightly adapted (see Proposition 9 and Corollary 7 in Appendix A for details). Finally, Proposition 3 is generalized by the next proposition, which

states that $p(\mathscr{S}) = p(\mathcal{H})$ and thus S and $p(\mathcal{H})$ are isomorphic (see the proof in Appendix I.2, Proposition 19 in [15]).

PROPOSITION 4 (Isometry between supplements). *$T^{1/2} : S \to \mathscr{S}$ is an isometry. Moreover,* $\operatorname{Im}(T^{1/2}) = p(\mathcal{H})$ *and $T^{1/2} : S \to p(\mathcal{H})$ is an isomorphism.*

We can also derive a version of Mercer's theorem, which does not make any more assumptions that are required for defining RKHSs. As we will not use it in this article, this proposition is only given in Appendix A.

Convergence results: In all convergence results stated below, *assumptions* (A1), (A2) *may be replaced by assumptions* (A1′), (A2′).

2.6. *Examples*. The property $\overline{\mathcal{H}} = S$ [where S is the orthogonal supplement of $\operatorname{Ker}(T)$], stated after Proposition 3, is important to understand what the space $\overline{\mathcal{H}}$ is, as we are minimizing over this closed and convex set. As a consequence the space $\mathcal{H}$ is dense in $L^2_{\rho_X}$ if and only if T is injective [or equivalently, $\operatorname{Ker}(T) = \{0\} \Leftrightarrow S = L^2_{\rho_X}$]. We detail below a few classical situations in which different configurations for the "inclusion" $\mathcal{H} \subset \overline{\mathcal{H}} \subset L^2_{\rho_X}$ appear:

1. *Finite-dimensional setting with linear kernel*: In finite dimension, with $\mathcal{X} = \mathbb{R}^d$ and $K(x, y) = x^\top y$, we have $\mathcal{H} = \mathbb{R}^d$, with the scalar product in $\langle u, v\rangle_{\mathcal{H}} = \sum_{i=1}^d u_i v_i$. This corresponds to usual parametric least-squares regression. If the support of ρ_X has a nonempty interior, then $\overline{\mathcal{H}} = \mathcal{H}$: $g_{\mathcal{H}}$ is the best linear estimator. Moreover, we have $\mathcal{H} = \overline{\mathcal{H}} \subsetneq L^2_{\rho_X}$: indeed $\operatorname{Ker}(T)$ is the set of functions such that $\mathbb{E}Xf(X) = 0$ (which is a large space).

2. *Translation-invariant kernels*: For instance, the Gaussian kernel over $\mathcal{X} = \mathbb{R}^d$, with X following a distribution with full support in $\mathbb{R}^d$: in such a situation we have $\mathcal{H} \subsetneq \overline{\mathcal{H}} = L^2_{\rho_X}$. This last equality holds more generally for all universal kernels, which include all kernels of the form $K(x, y) = q(x - y)$ where q has a summable strictly positive Fourier transform [27, 37]. These kernels are exactly the kernels such that T is an injective endomorphism of $L^2_{\rho_X}$.

3. *Splines over the circle*: When $X \sim \mathcal{U}[0; 1]$ and $\mathcal{H}$ is the set of m-times periodic weakly differentiable functions (see Section 5), we have in general $\mathcal{H} \subsetneq \overline{\mathcal{H}} \subsetneq L^2_{\rho_X}$. In such a case, $\ker(T) = \operatorname{span}(x \mapsto 1)$, and $\overline{\mathcal{H}} \oplus \operatorname{span}(x \mapsto 1) = L^2_{\rho_X}$, that is, we can approximate any zero-mean function.

Many examples and more details may be found in [2, 35, 42]. In particular, kernels on nonvectorial objects may be defined (e.g., sequences, graphs or measures).

2.7. *Convergence rates*. In order to be able to establish rates of convergence in this infinite-dimensional setting, we have to make assumptions on the objective function and on the covariance operator eigenvalues. In order to account for all cases (finite and infinite dimensions), we now consider eigenvalues ordered in *nonincreasing* order, that is, we assume that the set I is either $\{1, \dots, d\}$ if the underlying space is d-dimensional or $\mathbb{N}^*$ if the underlying space has infinite dimension.

(A3) We denote $(\mu_i)_{i\in I}$ the sequence of nonzero eigenvalues of the operator T, in decreasing order. We assume $\mu_i \le \frac{s^2}{i^\alpha}$ for some $\alpha > 1$ [so that $\operatorname{tr}(T) < \infty$], with $s \in \mathbb{R}_+$.

(A4) $g_{\mathcal{H}} \in T^r(L^2_{\rho_X})$ with $r \ge 0$, and as a consequence $\|T^{-r}(g_{\mathcal{H}})\|_{L^2_{\rho_X}} < \infty$.

We chose such assumptions in order to make the comparison with the existing literature as easy as possible, for example, [10, 47]. However, some other assumptions may be found as in [3, 20].

Dependence on α and r. The two parameters r and α intuitively parameterize the strengths of our assumptions:

- In assumption (A3), a bigger α makes the assumption stronger: it means the reproducing kernel Hilbert space is smaller, that is, if (A3) holds with some constant α, then it also holds for any $\alpha' < \alpha$. Moreover, if T is reduced in the Hilbertian basis $(\phi_i)_i$ of $L^2_{\rho_X}$, we have an effective search space $S = \{\sum_{i=1}^\infty b_i\phi_i / \sum_{i=1}^\infty \frac{b_i^2}{\mu_i} < \infty\}$: the smaller the eigenvalues, the smaller the space. Note that since $\operatorname{tr}(T)$ is finite, (A3) is always true for $\alpha = 1$.
- In assumption (A4), for a fixed α, a bigger r makes the assumption stronger, that is, the function $g_{\mathcal{H}}$ is actually smoother. Indeed, considering that (A4) may be rewritten $g_{\mathcal{H}} \in T^r(L^2_{\rho_X})$ and for any $r < r'$, $T^{r'}(L^2_{\rho_X}) \subset T^r(L^2_{\rho_X})$. In other words, $\{T^r(L^2_{\rho_X})\}_{r\ge 0}$ are decreasing (r growing) subspaces of $L^2_{\rho_X}$.

  For $r = 1/2$, $T^{1/2}(L^2_{\rho_X}) = \mathcal{H}$; moreover, for $r \ge 1/2$, our best approximation function $g_{\mathcal{H}} \in \overline{\mathcal{H}}$ is in fact in $\mathcal{H}$, that is, the optimization problem in the RKHS $\mathcal{H}$ is attained by a function of finite norm. However, for $r < 1/2$ it is not attained.
- Furthermore, it is worth pointing the stronger assumption which is often used in the finite dimensional context, namely $\operatorname{tr}(\Sigma^{1/\alpha}) = \sum_{i\in I} \mu_i^{1/\alpha}$ finite. It turns out that this is a stronger assumption, indeed, since we have assumed that the eigenvalues (μ_i) are arranged in nonincreasing order, if $\operatorname{tr}(\Sigma^{1/\alpha})$ is finite, then (A3) is satisfied for $s^2 = [2\operatorname{tr}(\Sigma^{1/\alpha})]^\alpha$. Such an assumption appears, for example, in Corollary 5.

Related assumptions. Assumptions (A3) and (A4) are adapted to our theoretical results, but some stricter assumptions are often used, that make comparison with existing work more direct. For comparison purposes, we will also use:

(a3) For any $i \in I = \mathbb{N}$, $u^2 \le i^\alpha \mu_i \le s^2$ for some $\alpha > 1$ and $u, s \in \mathbb{R}_+$.

(a4) We assume the coordinates $(v_i)_{i\in\mathbb{N}}$ of $g_{\mathcal{H}} \in L^2_{\rho_X}$ in the eigenbasis $(\phi_i)_{i\in\mathbb{N}}$ (for $\|\cdot\|_{L^2_{\rho_X}}$) of T are such that $v_i i^{\delta/2} \le W$, for some $\delta > 1$ and $W \in \mathbb{R}_+$ (so that $\|g_{\mathcal{H}}\|_{L^2_{\rho_X}} < \infty$).

Assumption (a3) directly imposes that the eigenvalues of T decay at rate $i^{-\alpha}$ (which imposes that there are infinitely many), and thus implies (A3). Together,

assumptions (a3) and (a4), imply assumptions (A3) and (A4), with any $\delta > 1 + 2\alpha r$. Indeed, we have

$$\|T^{-r} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} = \sum_{i\in\mathbb{N}} \nu_i^2 \mu_i^{-2r} \leq \frac{W^2}{u^{4r}} \sum_{i\in\mathbb{N}} i^{-\delta+2\alpha r},$$

which is finite for $2\alpha r - \delta < -1$. Thus, the supremum element of the set of r such that (A4) holds is such that $\delta = 1 + 2\alpha r$. Thus, when comparing assumptions (A3)–(A4) and (a3)–(a4), we will often make the identification above, that is, $\delta = 1 + 2\alpha r$.

The main advantage of the new assumptions is their interpretation when the basis $(\phi_i)_{i\in I}$ is common for several RKHSs (such as the Fourier basis for splines, see Section 5): (a4) describes the decrease of the coordinates of the best function $g_{\mathcal{H}} \in L^2_{\rho_X}$ *independently of the chosen RKHS*. Thus, the parameter δ characterizes the prediction function, while the parameter α characterizes the RKHS.

3. Stochastic approximation in Hilbert spaces. In this section, we consider estimating a prediction function $g \in \mathcal{H}$ from observed data, and we make the following assumption:

(A5) For $n \geq 1$, the random variables $(x_n, y_n) \in \mathcal{X} \times \mathbb{R}$ are independent and identically distributed with distribution ρ.

Our goal is to estimate a function $g \in \mathcal{H}$ from data, such that $\varepsilon(g) = \mathbb{E}(Y - g(X))^2$ is as small as possible. As shown in Section 2, this is equivalent to minimizing $\|g - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}$. The two main approaches to define an estimator is by regularization or by stochastic approximation (and combinations thereof). See also approaches by early-stopped gradient descent on the empirical risk in [46].

3.1. *Regularization and linear systems.* Given n observations, regularized empirical risk minimization corresponds to minimizing with respect to $g \in \mathcal{H}$ the following objective function:

$$\frac{1}{n}\sum_{i=1}^{n} \big(y_i - g(x_i)\big)^2 + \lambda \|g\|^2_{\mathcal{H}}.$$

Although the problem is formulated in a potentially infinite-dimensional Hilbert space, through the classical representer theorem [22, 33, 35], the unique (if $\lambda > 0$) optimal solution may be expressed as $\hat{g} = \sum_{i=1}^n a_i K_{x_i}$, and $a \in \mathbb{R}^n$ may be obtained by solving the linear system $(\mathbf{K} + \lambda I)a = \mathbf{y}$, where $\mathbf{K} \in \mathbb{R}^{n\times n}$ is the kernel matrix, a.k.a. the Gram matrix, composed of pairwise kernel evaluations $\mathbf{K}_{ij} = K(x_i, x_j)$, $i, j = 1, \ldots, n$, and $\mathbf{y}$ is the n-dimensional vector of all n responses y_i, $i = 1, \ldots, n$.

The running-time complexity to obtain $a \in \mathbb{R}^n$ is typically $O(n^3)$ if no assumptions are made, but several algorithms may be used to lower the complexity and

obtain an approximate solution, such as conjugate gradient [18] or column sampling (a.k.a. Nyström method) [3, 26, 44].

In terms of convergence rates, assumptions (a3)–(a4) allow to obtain convergence rates that decompose $\varepsilon(\hat{g}) - \varepsilon(g_{\mathcal{H}}) = \|\hat{g} - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}$ as the sum of two asymptotic terms [3, 10, 20]:

- *Variance term*: $O(\sigma^2 n^{-1} \lambda^{-1/\alpha})$, which is decreasing with λ, where σ^2 characterizes the noise variance, for example, in the homoscedastic case (i.i.d. additive noise), the marginal variance of the noise; see assumption (A6) for the detailed assumption that we need in our stochastic approximation context.
- *Bias term*: $O(\lambda^{\min\{(\delta-1)/\alpha, 2\}})$, which is increasing with λ. Note that the corresponding r from assumptions (A3)–(A4) is $r = (\delta - 1)/2\alpha$, and the bias term becomes proportional to $\lambda^{\min\{2r, 2\}}$.

There are then two regimes:

- *Optimal predictions*: If $r < 1$, then the optimal value of λ (that minimizes the sum of two terms and makes them asymptotically equivalent) is proportional to $n^{-\alpha/(2r\alpha+1)} = n^{-\alpha/\delta}$ and the excess prediction error $\|\hat{g} - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} = O(n^{-2\alpha r/(2\alpha r+1)}) = O(n^{-1+1/\delta})$, and the resulting procedure is then "optimal" in terms of estimation of $g_{\mathcal{H}}$ in $L^2_{\rho_X}$ (see Section 4 for details).
- *Saturation*: If $r \geq 1$, where the optimal value of λ (that minimizes the sum of two terms and makes them equivalent) is proportional to $n^{-\alpha/(2\alpha+1)}$, and the excess prediction error is less than $O(n^{-2\alpha/(2\alpha+1)})$, which is suboptimal. Although assumption (A4) is valid for a larger r, the rate is the same than if $r = 1$.

In this paper, we consider a stochastic approximation framework with improved running-time complexity and similar theoretical behavior than regularized empirical risk minimization, with the advantage of (a) needing a single pass through the data and (b) simple assumptions.

3.2. *Stochastic approximation.* Using the reproducing property, we have for any $g \in \mathcal{H}$, $\varepsilon(g) = \mathbb{E}(Y - g(X))^2 = \mathbb{E}(Y - \langle g, K_X \rangle_{\mathcal{H}})^2$, with gradient (defined with respect to the dot-product in $\mathcal{H}$) $\nabla\varepsilon(g) = -2\mathbb{E}[(Y - \langle g, K_X \rangle_{\mathcal{H}}) K_X]$.

Thus, for each pair of observations (x_n, y_n), we have $\nabla\varepsilon(g) = -2\mathbb{E}[(y_n - \langle g, K_{x_n} \rangle_{\mathcal{H}}) K_{x_n}]$, and thus, the quantity $[-(y_n - \langle g, K_{x_n} \rangle_{\mathcal{H}}) K_{x_n}] = [-(y_n - g(x_n)) K_{x_n}]$ is an *unbiased stochastic* (*half*) *gradient*. We thus consider the stochastic gradient recursion, in the Hilbert space $\mathcal{H}$, started from a function $g_0 \in \mathcal{H}$ (taken to be zero in the following):

$$g_n = g_{n-1} - \gamma_n \big[y_n - \langle g_{n-1}, K_{x_n} \rangle_{\mathcal{H}} \big] K_{x_n} = g_{n-1} - \gamma_n \big[y_n - g_{n-1}(x_n) \big] K_{x_n},$$

where γ_n is the *step-size*.

We may also apply the recursion using representants. Indeed, if $g_0 = 0$, which we now assume, then for any $n \geq 1$,

$$g_n = \sum_{i=1}^{n} a_i K_{x_i},$$

with the following recursion on the sequence $(a_n)_{n\geq 1}$:

$$a_n = -\gamma_n\big(g_{n-1}(x_n) - y_n\big) = -\gamma_n\left(\sum_{i=1}^{n-1} a_i K(x_n, x_i) - y_n\right).$$

We also output the averaged iterate defined as

$$\bar{g}_n = \frac{1}{n+1}\sum_{k=0}^{n} g_k = \frac{1}{n+1}\sum_{i=1}^{n}\left(\sum_{j=1}^{i} a_j\right) K_{x_i}. \tag{3.1}$$

Running-time complexity. The running time complexity is $O(i)$ for iteration i—if we assume that kernel evaluations are $O(1)$, and thus $O(n^2)$ after n steps. This is a serious limitation for practical applications. Several authors have considered expanding g_n on a subset of all (K_{x_i}), which allows to bring down the complexity of each iteration and obtain an overall linear complexity is n [8, 13], but this comes at the expense of not obtaining the sharp generalization errors that we obtain in this paper. Note that when studying regularized least-squares problem (i.e., adding a penalisation term), one has to update every coefficient $(a_i)_{1\leq i\leq n}$ at step n, while in our situation, only a_n is computed at step n.

Relationship to previous works. Similar algorithms have been studied before [23, 32, 45, 47, 48], under various forms. Especially, in [23, 39, 45, 48] a regularization term is added to the loss function [thus considering the following problem: $\arg\min_{f\in\mathcal{H}} \varepsilon(f) + \lambda\|f\|_K^2$]. In [32, 47], neither regularization nor averaging procedure are considered, but in the second case, multiple pass through the data are considered. In [48], a nonregularized averaged procedure equivalent to ours is considered. However, the step-sizes γ_n which are proposed, as well as the corresponding analysis, are different. Our step-sizes are larger and our analysis uses more directly the underlying linear algebra to obtain better rates (while the proof of [48] is applicable to all smooth losses).

Step-sizes. We are mainly interested in two different types of step-sizes (a.k.a. *learning rates*): the sequence $(\gamma_i)_{1\leq i\leq n}$ may be either:

1. A subsequence of a universal sequence $(\gamma_i)_{i\in\mathbb{N}}$, we refer to this situation as the "*online setting*." Our bounds then hold for any of the iterates.

2. A sequence of the type $\gamma_i = \Gamma(n)$ for $i \leq n$, which will be referred to as the "*finite horizon setting*:" in this situation the number of samples is assumed to be known and fixed and we chose a constant step-size which may depend on this number. Our bounds then hold only for the last iterate.

In practice, it is important to have an online procedure, to be able to deal with huge amounts of data (potentially infinite). However, the analysis is easier in the "finite horizon" setting. Some *doubling tricks* allow to pass to varying steps [19], but it may not be not fully satisfactory in practice as it creates jumps at every n which is a power of two.

3.3. *Extra regularity assumptions.* We denote by $\Xi = (Y - g_{\mathcal{H}}(X))K_X$ the residual, a random element of $\mathcal{H}$. We have $\mathbb{E}[\Xi] = 0$ but in general we do not have $\mathbb{E}[\Xi|X] = 0$ (unless the model of homoscedastic regression is well specified). We make the following extra assumption:

(A6) There exists $\sigma > 0$ such that $\mathbb{E}[\Xi \otimes \Xi] \preccurlyeq \sigma^2 \Sigma$, where $\preccurlyeq$ denotes the order between self-adjoint operators.

In other words, for any $f \in \mathcal{H}$, we have $\mathbb{E}[(Y - g_{\mathcal{H}}(X))^2 f(X)^2] \leq \sigma^2 \mathbb{E}[f(X)^2]$.

In the well-specified homoscedastic case, we have that $(Y - g_{\mathcal{H}}(X))$ is independent of X and with $\sigma^2 = \mathbb{E}[(Y - g_{\mathcal{H}}(X))^2]$, $\mathbb{E}[\Xi|X] = \sigma^2 \Sigma$ is clear: the constant σ^2 in the first part of our assumption characterizes the noise amplitude. Moreover, when $|Y - g_{\mathcal{H}}(X)|$ is a.s. bounded by σ^2, we have (A6).

We first present the results in the *finite horizon* setting in Section 3.4 before turning to the *online* setting in Section 3.5.

3.4. *Main results* (*finite horizon*). We can first get some guarantee on the consistency of our estimator, for any small enough constant step-size.

THEOREM 1. *Assume* (A1)–(A6), *then for any constant choice* $\gamma_n = \gamma_0 < \frac{1}{2R^2}$, *the prediction error of* $\bar{g}_n$ *converges to the one of* $g_{\mathcal{H}}$, *that is*,

$$\mathbb{E}\big[\varepsilon(\bar{g}_n) - \varepsilon(g_{\mathcal{H}})\big] = \mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \overset{n\to\infty}{\longrightarrow} 0. \tag{3.2}$$

The expectation is considered with respect to the distribution of the sample $(x_i, y_i)_{1\leq i\leq n}$, as in all the following theorems (note that $\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}$ is itself a different expectation with respect to the law ρ_X).

Theorem 1 means that for the simplest choice of the learning rate as a constant, our estimator tends to the perform as well as the best estimator in the class $\mathcal{H}$. Note that in general, the convergence in $\mathcal{H}$ is meaningless if $r < 1/2$. The following results will state some assertions on the speed of such a convergence; our main result, in terms of generality is the following.

THEOREM 2 (Complete bound, γ constant, finite horizon). *Assume* (A1)–(A6) *and* $\gamma_i = \gamma = \Gamma(n)$, *for* $1 \leq i \leq n$. *If* $\gamma R^2 \leq 1/4$:

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \leq \frac{4\sigma^2}{n}\big(1 + (s^2\gamma n)^{1/\alpha}\big) + 4(1 + \tau_{n,\gamma,r,\alpha})\frac{\|T^{-r} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}}{\gamma^{2r} n^{2\min\{r,1\}}},$$

where $\tau_{n,\gamma,r,\alpha} := (R^{2\alpha}\gamma^{1+\alpha} n s^2)^{\frac{2r-1}{\alpha}}$ *if* $r \geq \frac{1}{2}$ *and* $\tau_{n,\gamma,r,\alpha} := 0$ *otherwise.*

We can make the following observations.

– *Proof*: Theorem 1 is directly derived from Theorem 2, which is proved in Appendix II.3 in [15]: we derive for our algorithm a new error decomposition and bound the different sources of error via algebraic calculations. More precisely, following the proof in Euclidean space [5], we first analyze (in Appendix II.2 in [15]) a closely related recursion (we replace $K_{x_n} \otimes K_{x_n}$ by its expectation Σ, and we thus refer to it as a semi-stochastic version of our algorithm):

$$g_n = g_{n-1} - \gamma_n(y_n K_{x_n} - \Sigma g_{n-1}).$$

It (a) leads to an easy computation of the main bias/variance terms of our result, (b) will be used to derive our main result by bounding the drifts between our algorithm and its semi-stochastic version. A more detailed sketch of the proof is given in Appendix B.

– *Bias/variance interpretation*: The two main terms have a simple interpretation. The first one is a variance term, which shows the effect of the noise σ^2 on the error. It is bigger when σ gets bigger and, moreover, it also gets bigger when γ is growing (bigger steps mean more variance). As for the second term, it is a bias term, which accounts for the distance of the initial choice (the null function in general) to the objective function. As a consequence, it is smaller when we make bigger steps.

– *Assumption* (A4): Our assumption (A4) for $r > 1$ is stronger than for $r = 1$ but we do not improve the bound. Indeed the bias term (see comments below) cannot decrease faster than $O(n^{-2})$: this phenomenon in known as saturation [16]. To improve our results with $r > 1$ it may be interesting to consider another type of averaging. In the following, $r < 1$ shall be considered as the main and most interesting case.

– *Relationship to regularized empirical risk minimization*: Our bound ends up being very similar to bounds for regularized empirical risk minimization, with the identification $\lambda = \frac{1}{\gamma n}$. It is thus no surprise that once we optimize for the value of γ, we recover the same rates of convergence. Note that in order to obtain convergence, we require that the step-size γ is bounded, which corresponds to an equivalent λ which has to be lower-bounded by $1/n$.

– *Residual term*: $\tau_{n,\gamma,r,\alpha}$ is a residual quantity which only appears when the prediction function $g_{\mathcal{H}}$ is in the RKHS, and will eventually be negligible for the choice of γ proposed in Corollary 2. Note that it also depends on s and R, a dependency hidden in notation $\tau_{n,\gamma,r,\alpha}$ to simplify a bit the notation.

– *Finite horizon*: Once again, this theorem holds in the finite horizon setting. That is, we first choose the number of samples we are going to use, then the learning rate as a constant. It allows us to choose γ as a function of n, in order to balance the main terms in the error bound. The trade-off must be understood as follows: a bigger γ increases the effect of the noise, but a smaller one makes it harder to forget the initial condition.

We may now deduce the following corollaries, with specific optimized values of γ.

COROLLARY 2 (Optimal constant γ). *Assume* (A1)–(A6) *and a constant step-size* $\gamma_i = \gamma = \Gamma(n)$, *for* $1 \leq i \leq n$:

1. *If* $\frac{\alpha-1}{2\alpha} < r$ *and* $\Gamma(n) = \gamma_0 n^{(-2\alpha\min\{r,1\}-1+\alpha)/(2\alpha\min\{r,1\}+1)}$, $\gamma_0 R^2 \leq 1/4$, *we have*:

$$\mathbb{E}\big(\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\big) \leq An^{-(2\alpha\min\{r,1\})/(2\alpha\min\{r,1\}+1)}, \tag{3.3}$$

with $A = 4(1 + (\gamma_0 s^2)^{1/\alpha})\sigma^2 + \frac{4(1+o(1))}{\gamma_0^{2r}}\|L_K^{-r} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}$.

2. *If* $0 < r < \frac{\alpha-1}{2\alpha}$, *with* $\Gamma(n) = \gamma_0$ *is constant*, $\gamma_0 R^2 \leq 1/4$, *we have*

$$\mathbb{E}\big(\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\big) \leq An^{-2r}, \tag{3.4}$$

with the same constant A.

We can make the following observations.

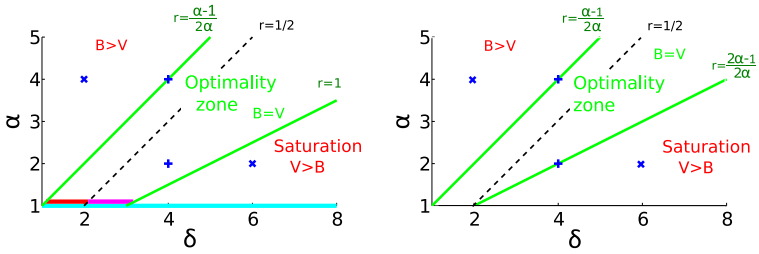
– *Limit conditions*: Assumption (A4), gives us some kind of "position" of the objective function with respect to our reproducing kernel Hilbert space. If $r \geq 1/2$, then $g_{\mathcal{H}} \in \mathcal{H}$. That means the regression function truly lies in the space in which we are looking for an approximation. However, it is not necessary neither to get the convergence result, which stands for any $r > 0$, nor to get the optimal rate (see definition in Section 4.2), which is also true for $\frac{\alpha-1}{2\alpha} < r < 1$.

– The quantity $o(1)$ in Equation (3.3) stands for $(\gamma_0 s^2 n^{-2\alpha^2 r+1})^{\frac{2r-1}{\alpha}}$ if $r \geq 1/2$ (0 otherwise) and is a quantity which decays to 0.

– *Evolution with* r *and* α: As it has been noticed above, a bigger α or r would be a stronger assumption. It is thus natural to get a rate which improves with a bigger α or r: the function $(\alpha, r) \mapsto \frac{2\alpha r}{2\alpha r+1}$ is increasing in both parameters.

– *Different regions*: in Figure 1(a), we plot in the plan of coordinates α, δ (with $\delta = 2\alpha r + 1$) our limit conditions concerning our assumptions, that is, $r = 1 \Leftrightarrow \delta = 2\alpha + 1$ and $\frac{\alpha-1}{2\alpha} = r \Leftrightarrow \alpha = \delta$. The region between the two green lines is the region for which the optimal rate of estimation is reached. The dashed lines stands for $r = 1/2$, which has appeared to be meaningless in our context.

The region $\alpha \geq \delta \Leftrightarrow \frac{\alpha-1}{2\alpha} > r$ corresponds to a situation where regularized empirical risk minimization would still be optimal, but with a regularization parameter λ that decays faster than $1/n$, and thus, our corresponding step-size



(a) Finite Horizon (b) Online

FIG. 1. *Behavior of convergence rates*: (*left*) *finite horizon and* (*right*) *online setting. We describe in the* (α, δ) *plan* (*with* $\delta = 2\alpha r + 1$) *the different optimality regions*: *between the two green lines*, *we achieve the optimal rate. On the left plot*, *the red* (*resp.*, *magenta and cyan*) *lines are the regions for which Zhang* [48] (*resp.*, *Yao and Tarrès* [39] *and Ying and Pontil* [47]) *proved to achieve the overall optimal rate* (*which may only be the case if* $\alpha = 1$). *The four blue points match the coordinates of the four couples* (α, δ) *that will be used in our simulations*: *they are spread over the different optimality regions.*

$\gamma = 1/(n\lambda)$ would not be bounded as a function of n. We thus saturate our step-size to a constant and the generalization error is dominated by the bias term.

The region $\alpha \leq (\delta - 1)/2 \Leftrightarrow r > 1$ corresponds to a situation where regularized empirical risk minimization reaches a saturating behavior. In our stochastic approximation context, the variance term dominates.

3.5. *Online setting.* We now consider the second case when the sequence of step-sizes does not depend on the number of samples we want to use (online setting).

The computations are more tedious in such a situation so that we will only state asymptotic theorems in order to understand the similarities and differences between the finite horizon setting and the online setting, especially in terms of limit conditions.

THEOREM 3 [Complete bound, $(\gamma_n)_n$ online]. *Assume* (A1)–(A6), *assume for any* i, $\gamma_i = \frac{\gamma_0}{i^\zeta}$, $\gamma_0 R^2 \leq 1/2$:

– *If* $0 < r(1-\zeta) < 1$, *if* $0 < \zeta < \frac{1}{2}$ *then*

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \leq O\left(\frac{\sigma^2 (s^2\gamma_n)^{1/\alpha}}{n^{1-1/\alpha}}\right) + O\left(\frac{\|L_K^{-r} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}}{(n\gamma_n)^{2r}}\right). \tag{3.5}$$

– *If* $0 < r(1-\zeta) < 1$, $\frac{1}{2} < \zeta$

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \leq O\left(\frac{\sigma^2 (s^2\gamma_n)^{1/\alpha}}{n^{1-1/\alpha}} \frac{1}{n\gamma_n^2}\right) + O\left(\frac{\|L_K^{-r} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}}{(n\gamma_n)^{2r}}\right). \tag{3.6}$$

The constant in the $O(\cdot)$ *notation only depend on* γ_0 *and* α.

Theorem 3 is proved in Appendix II.4 in [15]. In the first case, the main bias and variance terms are the same as in the finite horizon setting, and so is the optimal choice of ζ. However, in the second case, the variance term behavior changes: it does not decrease any more when ζ increases beyond $1/2$. Indeed, in such a case our constant averaging procedure puts to much weight on the first iterates, thus we do not improve the variance bound by making the learning rate decrease faster. Other types of averaging, as proposed, for example, in [25], could help to improve the bound.

Moreover, this extra condition thus changes a bit the regions where we get the optimal rate [see Figure 1(b)], and we have the following corollary.

COROLLARY 3 (Optimal decreasing γ_n). *Assume* (A1)–(A6) [*in this corollary,* $O(\cdot)$ *stands for a constant depending on* α, $\|L_K^{-r} g_{\mathcal{H}}\|_{L^2_{\rho_X}}$, s, σ^2, γ_0 *and universal constants*]:

1. *If* $\frac{\alpha-1}{2\alpha} < r < \frac{2\alpha-1}{2\alpha}$, *with* $\gamma_n = \gamma_0 n^{(-2\alpha r-1+\alpha)/(2\alpha r+1)}$ *for any* $n \geq 1$ *we get the rate*

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} = O\big(n^{-2\alpha r/(2\alpha r+1)}\big). \tag{3.7}$$

2. *If* $\frac{2\alpha-1}{2\alpha} < r$, *with* $\gamma_n = \gamma_0 n^{-1/2}$ *for any* $n \geq 1$, *we get the rate*

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} = O\big(n^{-(2\alpha-1)/(2\alpha)}\big). \tag{3.8}$$

3. *If* $0 < r < \frac{\alpha-1}{2\alpha}$, *with* $\gamma_n = \gamma_0$ *for any* $n \geq 1$, *we get the rate given in* (3.4). *Indeed the choice of a constant learning rate naturally results in an online procedure.*

This corollary is directly derived from Theorem 3, balancing the two main terms. The only difference with the finite horizon setting is the shrinkage of the optimality region as the condition $r < 1$ is replaced by $r < \frac{2\alpha-1}{2\alpha} < 1$ [see Figure 1(b)]. In the next section, we relate our results to existing work.

4. Links with existing results. In this section, we relate our results from the previous section to existing results.

4.1. *Euclidean spaces.* Recently, Bach and Moulines showed in [5] that for least squares regression, averaged stochastic gradient descent achieved a rate of $O(1/n)$, in a finite-dimensional Hilbert space (Euclidean space), under the same assumptions as above (except the first one of course), which is replaced by:

(A1-f) $\mathcal{H}$ is a d-dimensional Euclidean space.

They showed the following result.

PROPOSITION 5 (Finite-dimensions [5]). *Assume* (A1-f), (A2)–(A6). *Then for* $\gamma = \frac{1}{4R^2}$,

$$\mathbb{E}\big[\varepsilon(\bar{g}_n) - \varepsilon(g_{\mathcal{H}})\big] \leq \frac{4}{n}\big[\sigma\sqrt{d} + R\|g_{\mathcal{H}}\|_{\mathcal{H}}\big]^2. \tag{4.1}$$

We show that we can deduce such a result from Theorem 2 (and even with comparable constants). Indeed under (A1-f) we have:

- If $\mathbb{E}[\|x_n\|^2] \leq R^2$ then $\Sigma \preccurlyeq R^2 I$ and (A3) is true for any $\alpha \geq 1$ with $s^2 = R^2 d^\alpha$. Indeed $\lambda_i \leq R^2$ if $i \leq d$ and $\lambda_i = 0$ if $i > d+1$ so that for any $\alpha > 1, i \in \mathbb{N}^*$, $\lambda_i \leq R^2 \frac{d^\alpha}{i^\alpha}$.
- As we are in a finite-dimensional space (A4) is true for $r = 1/2$ as $\|T^{-1/2} \times g_{\mathcal{H}}\|^2_{\mathcal{L}^2_{\rho_X}} = \|g_{\mathcal{H}}\|^2_{\mathcal{H}}$.

Under such remarks, the following corollary may be deduced from Theorem 2.

COROLLARY 4. *Assume* (A1-f), (A2)–(A6), *then for any* $\alpha > 1$, *with* $\gamma R^2 \leq 1/4$,

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \leq \frac{4\sigma^2}{n}\big(1 + \big(R^2\gamma d^\alpha n\big)^{1/\alpha}\big) + 4\frac{\|g_{\mathcal{H}}\|^2_{\mathcal{H}}}{n\gamma}$$

so that, *when* $\alpha \to \infty$,

$$\mathbb{E}\big[\varepsilon(\bar{g}_n) - \varepsilon(g_{\mathcal{H}})\big] \leq \frac{4}{n}\left(\sigma\sqrt{d} + R\|g_{\mathcal{H}}\|_{\mathcal{H}}\frac{1}{\sqrt{\gamma R^2}}\right)^2.$$

This bound is easily comparable to (4.1) and shows that our more general analysis has not lost too much. Moreover, our learning rate is proportional to $n^{-1/(2\alpha+1)}$ with $r = 1/2$, so tends to behave like a constant when $\alpha \to \infty$, which recovers the constant step set-up from [5].

Moreover, N. Flammarion proved (personnal communication, 05/2014), using the same type of techniques that their bound could be extended to

$$\mathbb{E}\big[\varepsilon(\bar{g}_n) - \varepsilon(g_{\mathcal{H}})\big] \leq 8\frac{\sigma^2 d}{n} + 4R^4(1 + \gamma d R^2)\frac{\|\Sigma^{-1/2} g_{\mathcal{H}}\|^2}{(\gamma R^2)^2 n^2}, \tag{4.2}$$

a result that may be deduced of the following more general corollaries of our Theorem 2.

COROLLARY 5. *Assume* (A1-f), (A2)–(A6), *and*, *for some* $q \in [-1/2; 1/2]$, $\|\Sigma^{-q} g_{\mathcal{H}}\|^2_{\mathcal{H}} = \|\Sigma^{-(q+1/2)} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} < \infty$, *then*

$$\begin{aligned}&\mathbb{E}\big[\varepsilon(\bar{g}_n) - \varepsilon(g_{\mathcal{H}})\big] \\ &\quad\leq 16\frac{\sigma^2 \operatorname{tr}(\Sigma^{1/\alpha})(\gamma n)^{1/\alpha}}{n} + 8\frac{R^{4(q+1/2)}(1 + \tau_{n,\gamma,q+1/2,\alpha})\|\Sigma^{-q} g_{\mathcal{H}}\|^2_{\mathcal{H}}}{(n\gamma R^2)^{2(q+1/2)}}.\end{aligned}$$

Such a result is derived from Theorem 2 and with the stronger assumption $\operatorname{tr}(\Sigma^{1/\alpha}) < \infty$ clearly satisfied in finite dimension, and with $r = q + 1/2$. Note that the result above is true for all values of $\alpha \geq 1$ and all $q \in [-1/2; 1/2]$ (for the ones with infinite $\|\Sigma^{-(q+1/2)} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}$, the statement is trivial). This shows that we may take the infimum over all possible $\alpha \leq 1$ and $q \in [-1/2; 1/2]$, showing adaptivity of the estimator to the spectral decay of Σ and the smoothness of the optimal prediction function $g_{\mathcal{H}}$.

The residual term $\tau_{n,\gamma,q+\frac{1}{2},\alpha}$ is the same as in Theorem 2. When $\alpha \to \infty$, it goes to a term which scales like $(\gamma d)^{2q}$, if $q \in [0; 1/2]$ (it is 0 otherwise).

Thus, with $\alpha \to \infty$, we obtain the following.

COROLLARY 6. *Assume* (A1-f), (A2)–(A6) *and for some* $q \in [-1/2; 1/2]$, $\|\Sigma^{-q} g_{\mathcal{H}}\|^2_{\mathcal{H}} = \|\Sigma^{-(q+1/2)} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} < \infty$, *then*

$$\mathbb{E}\big[\varepsilon(\bar{g}_n) - \varepsilon(g_{\mathcal{H}})\big] \leq 16\frac{\sigma^2 d}{n} + 8R^{4(q+1/2)}\big(1 + (\gamma R^2 d)^{2q \vee 0}\big)\frac{\|\Sigma^{-q} g_{\mathcal{H}}\|^2_{\mathcal{H}}}{(n\gamma R^2)^{2(q+1/2)}}.$$

- This final result bridges the gap between Proposition 5 ($q = 0$), and its extension (4.2) ($q = 1/2$). The constants 16 and 8 come from the upper bounds $(a+b)^2 \leq a^2 + b^2$ and $1 + 1/\sqrt{d} \leq 2$ and are thus nonoptimal.
- Moreover, we can also derive from Corollary 5, with $\alpha = 1$, $q = 0$, and $\gamma \propto n^{-1/2}$, we recover the rate $O(n^{-1/2})$ (where the constant does not depend on the dimension d of the Euclidean space). Such a rate was described, for example, in [29].

Note that linking our work to the finite-dimensional setting is made using the fact that our assumption (A3) is true for any $\alpha > 1$.

4.2. *Optimal rates of estimation.* In some situations, our stochastic approximation framework leads to "optimal" rates of prediction in the following sense. In [10], Theorem 2, a minimax lower bound was given: let $\mathcal{P}(\alpha, r)$ ($\alpha > 1, r \in [1/2, 1]$) be the set of all probability measures ρ on $\mathcal{X} \times \mathcal{Y}$, such that:

- $|y| \leq M_\rho$ almost surely,
- $T^{-r} g_\rho \in L^2_{\rho_X}$,
- the eigenvalues $(\mu_j)_{j \in \mathbb{N}}$ arranged in a nonincreasing order, are subject to the decay $\mu_j = O(j^{-\alpha})$.

Then the following minimax lower rate stands:

$$\liminf_{n\to\infty} \inf_{g_n} \sup_{\rho \in \mathcal{P}(b,r)} \mathbb{P}\big\{\varepsilon(g_n) - \varepsilon(g_\rho) > Cn^{-2r\alpha/(2r\alpha+1)}\big\} = 1,$$

for some constant $C > 0$ where the infimum in the middle is taken over all algorithms that are maps $((x_i, y_i)_{1 \leq i \leq n}) \mapsto g_n \in \mathcal{H}$.

When making assumptions (a3)–(a4), the assumptions regarding the prediction problem (i.e., the optimal function g_ρ) are summarized in the decay of the components of g_ρ in an orthonormal basis, characterized by the constant δ. Here, the minimax rate of estimation (see, e.g., [21]) is $O(n^{-1+1/\delta})$ which is the same as $O(n^{-2r\alpha/(2r\alpha+1)})$ with the identification $\delta = 2\alpha r + 1$.

That means the rate we get is optimal for $\frac{\alpha-1}{2\alpha} < r < 1$ in the finite horizon setting, and for $\frac{\alpha-1}{2\alpha} < r < \frac{2\alpha-1}{2\alpha}$ in the online setting. This is the region between the two green lines on Figure 1.

4.3. *Regularized stochastic approximation.* It is interesting to link our results to what has been done in [45] and [39] in the case of regularized least-mean-squares, so that the recursion is written

$$g_n = g_{n-1} - \gamma_n\big((g_{n-1}(x_n) - y_n)K_{x_n} + \lambda_n g_{n-1}\big)$$

with $(g_{n-1}(x_n) - y_n)K_{x_n} + \lambda_n g_{n-1}$ an unbiased gradient of $\frac{1}{2}\mathbb{E}_\rho[(g(x) - y)^2] + \frac{\lambda_n}{2}\|g\|^2$. In [39], the following result is proved (*Remark* 2.8 following *Theorem* C).

THEOREM 4 (Regularized, nonaveraged stochastic gradient [39]). *Assume that $T^{-r}g_\rho \in L^2_{\rho_X}$ for some $r \in [1/2, 1]$. Assume the kernel is bounded and $\mathcal{Y}$ compact. Then with probability at least $1 - \kappa$, for all $t \in \mathbb{N}$,*

$$\varepsilon(g_n) - \varepsilon(g_\rho) \leq O_\kappa\big(n^{-2r/(2r+1)}\big),$$

where O_κ stands for a constant which depends on κ.

No assumption is made on the covariance operator beyond being trace class, but only on $\|T^{-r}g_\rho\|_{L^2_{\rho_X}}$ [thus no assumption (A3)]. A few remarks may be made:

1. They get almost-sure convergence, when we only get convergence in expectation. We could perhaps derive a.s. convergence by considering moment bounds in order to be able to derive convergence in high probability and to use Borel–Cantelli lemma.
2. They only assume $\frac{1}{2} \leq r \leq 1$, which means that they assume the regression function to lie in the RKHS.

4.4. *Unregularized stochastic approximation.* In [47], Ying and Pontil studied the same unregularized problem as we consider, under assumption (A4). They obtain the same rates as above [$n^{-2r/(2r+1)}\log(n)$] in both online case (with $0 \leq r \leq \frac{1}{2}$) and finite horizon setting ($0 < r$).

They led as an open problem to improve bounds with some additional information on some decay of the eigenvalues of T, a question which is answered here.

Moreover, Zhang [48] also studies stochastic gradient descent algorithms in an unregularized setting, also with averaging. As described in [47], his result

is stated in the linear kernel setting but may be extended to kernels satisfying $\sup_{x\in\mathcal{X}} K(x,x) \leq R^2$. Ying and Pontil derive from Theorem 5.2 in [48] the following proposition.

PROPOSITION 6 (Short step-sizes [48]). *Assume we consider the algorithm defined in Section 3.2 and output $\bar{g}_n$ defined by (3.1). Assume the kernel K satisfies $\sup_{x\in\mathcal{X}} K(x,x) \leq R^2$. Finally, assume g_ρ satisfies assumption* (A4) *with $0 < r < 1/2$. Then in the finite horizon setting, with $\Gamma(n) = \frac{1}{4R^2} n^{-2r/(2r+1)}$, we have*

$$\mathbb{E}\big[\varepsilon(\bar{g}_n) - \varepsilon(g_{\mathcal{H}})\big] = O\big(n^{-2r/(2r+1)}\big).$$

Moreover, note that we may derive their result from Corollary 2. Indeed, using $\Gamma(n) = \gamma_0 n^{-2r/(2r+1)}$, we get a bias term which is of order $n^{-2r/(2r+1)}$ and a variance term of order $n^{-1+1/(2r\alpha+\alpha)}$ which is smaller. Our analysis thus recovers their convergence rate with their step-size. Note that this step-size is significantly smaller than ours, and that the resulting bound is worse (but their result holds in more general settings than least-squares). See more details in Section 4.5.

4.5. *Summary of results.* All three algorithms are variants of the following:

$$\begin{aligned} &g_0 = 0, \\ \forall n \geq 1, \qquad &g_n = (1-\lambda_n) g_{n-1} - \gamma_n \big(y_n - g_{n-1}(x_n)\big) K_{x_n}. \end{aligned}$$

But they are studied under different settings, concerning regularization, averaging, assumptions: we sum up in Table 1 the settings of each of these studies. For each of them, we consider the finite horizon settings, where results are generally better.

We can make the following observations.

TABLE 1

Summary of assumptions and results (step-sizes, rates and conditions) for our three regions of convergence and related approaches. We focus on finite-horizon results

Algorithm type	**Ass. (A3)**	**Ass. (A4)**	γ_n	λ_n	**Rate**	**Conditions**
This paper	yes	yes	1	0	n^{-2r}	$r < \frac{\alpha-1}{2\alpha}$
This paper	yes	yes	$n^{-(2\alpha r+1-\alpha)/(2\alpha r+1)}$	0	$n^{-2\alpha r/(2\alpha r+1)}$	$\frac{\alpha-1}{2\alpha} < r < 1$
This paper	yes	yes	$n^{-(\alpha+1)/(2\alpha+1)}$	0	$n^{-2\alpha/(2\alpha+1)}$	$r > 1$
Zhang [48]	no	yes	$n^{-2r/(2r+1)}$	0	$n^{-2r/(2r+1)}$	$0 \leq r \leq \frac{1}{2}$
Tarrès & Yao [39]	no	yes	$n^{-2r/(2r+1)}$	$n^{-1/(2r+1)}$	$n^{-2r/(2r+1)}$	$\frac{1}{2} \leq r \leq 1$
Ying & Pontil [47]	no	yes	$n^{-2r/(2r+1)}$	0	$n^{-2r/(2r+1)}$	$r > 0$

- *Dependence of the convergence rate on* α: For learning with any kernel with $\alpha > 1$, we strictly improve the asymptotic rate compared to related methods that only assume summability of eigenvalues: indeed, the function $x \mapsto x/(x+1)$ is increasing on $\mathbb{R}^+$. If we consider a given optimal prediction function and a given kernel with which we are going to learn the function, considering the decrease in eigenvalues allows to adapt the step-size and obtain an improved learning rate. Namely, we improved the previous rate $\frac{-2r}{2r+1}$ up to $\frac{-2\alpha r}{2\alpha r+1}$.
- *Worst-case result in* r: In the setting of assumptions (a3), (a4), given δ, the optimal rate of convergence is known to be $O(n^{-1+1/\delta})$, where $\delta = 2\alpha r + 1$. We thus get the optimal rate, as soon as $\alpha < \delta < 2\alpha + 1$, while the other algorithms get the sub-optimal rate $n^{(\delta-1)/(\delta+\alpha-1)}$ under various conditions. Note that this sub-optimal rate becomes close to the optimal rate when α is close to one, that is, in the *worst-case* situation. Thus, in the worst case (α arbitrarily close to one), all methods behave similarly, but for any particular instance where $\alpha > 1$, our rates are better.
- Choice of kernel: In the setting of assumptions (a3), (a4), given δ, in order to get the optimal rate, we may choose the kernel (i.e., α) such that $\alpha < \delta < 2\alpha + 1$ (i.e., neither too big, nor too small), while other methods need to choose a kernel for which α is as close to one as possible, which may not be possible in practice.
- *Improved bounds*: Ying and Pontil [47] only give asymptotic bounds, while we have exact constants for the finite horizon case. Moreover, there are some logarithmic terms in [47] which disappear in our analysis.
- *Saturation*: Our method does saturate for $r > 1$, while the nonaveraged framework of [47] does not (but does not depend on the value of α). We conjecture that a proper nonuniform averaging scheme (that puts more weight on the latest iterates), we should get the best of both worlds.

5. Experiments on artificial data. Following [47], we consider synthetic examples with smoothing splines on the circle, where our assumptions (A3)–(A4) are easily satisfied.

5.1. *Splines on the circle.* The simplest example to match our assumptions may be found in [43]. We consider $Y = g_\rho(X) + \varepsilon$, with $X \sim \mathcal{U}[0; 1]$ is a uniform random variable in $[0, 1]$, and g_ρ in a particular RKHS (which is actually a Sobolev space).

Let $\mathcal{H}$ be the collection of all zero-mean periodic functions on $[0; 1]$ of the form

$$f : t \mapsto \sqrt{2}\sum_{i=1}^{\infty} a_i(f)\cos(2\pi i t) + \sqrt{2}\sum_{i=1}^{\infty} b_i(f)\sin(2\pi i t),$$

with

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty}\bigl(a_i(f)^2 + b_i(f)^2\bigr)(2\pi i)^{2m} < \infty.$$

This means that the mth derivative of f, $f^{(m)}$ is in $\mathcal{L}^2([0;1])$. We consider the inner product

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} (2\pi i)^{2m} \big(a_i(f)a_i(g) + b_i(f)b_i(g)\big).$$

It is known that $\mathcal{H}$ is an RKHS and that the reproducing kernel $R_m(s,t)$ for $\mathcal{H}$ is

$$\begin{aligned} R_m(s,t) &= \sum_{i=1}^{\infty} \frac{2}{(2\pi i)^{2m}} \big[\cos(2\pi i s)\cos(2\pi i t) + \sin(2\pi i s)\sin(2\pi i t)\big] \\ &= \sum_{i=1}^{\infty} \frac{2}{(2\pi i)^{2m}} \cos\big(2\pi i (s-t)\big). \end{aligned}$$

Moreover, the study of Bernoulli polynomials gives a close formula for $R(s,t)$, that is,

$$R_m(s,t) = \frac{(-1)^{m-1}}{(2m)!} B_{2m}\big(\{s-t\}\big),$$

with B_m denoting the mth Bernoulli polynomial and $\{s-t\}$ the fractional part of $s-t$ [43].

We can derive the following proposition for the covariance operator which means that our assumption (A3) is satisfied for our algorithm in $\mathcal{H}$ when $X \sim \mathcal{U}[0;1]$, with $\alpha = 2m$, and $s = 2(1/2\pi)^m$.

PROPOSITION 7 (Covariance operator for smoothing splines). *If $X \sim \mathcal{U}[0;1]$, then in $\mathcal{H}$:*

1. *The eigenvalues of Σ are all of multiplicity 2 and are $\lambda_i = (2\pi i)^{-2m}$.*
2. *The eigenfunctions are $\phi_i^c : t \mapsto \sqrt{2}\cos(2\pi i t)$ and $\phi_i^s : t \mapsto \sqrt{2}\sin(2\pi i t)$.*

PROOF. For ϕ_i^c, we have (a similar argument holds for ϕ_i^s)

$$\begin{aligned} T\big(\phi_i^c\big)(s) &= \int_0^1 R(s,t)\sqrt{2}\cos(2\pi i t)\,dt \\ &= \left(\int_0^1 \frac{2}{(2i\pi)^{2m}} \sqrt{2}\cos(2\pi i t)^2\,dt\right)\cos(2\pi i s) = \lambda_i \sqrt{2}\cos(2\pi i s) \\ &= \lambda_i \phi_i^c(s). \end{aligned}$$

It is well knog0 10.9589 1770.499[(i)3o

Finally, considering $g(x) = B_{r/2}(x)$ with $\alpha = 2r + 1 \in 2\mathbb{N}$, our assumption (A4) holds. Indeed it implies (a3)–(a4), with $\beta > 1$, $\alpha = 2r + 1$, since for any $k \in \mathbb{N}$, $B_k(x) = (-1)^{k+1} 2k! \sum_{i=1}^{\infty} \frac{\cos(2i\pi x - k\pi/2)}{(2i\pi)^k}$ (see, e.g., [1]).

We may notice a few points:

1. Here the eigenvectors do not depend on the kernel choice, only the renormalisation constant depends on the choice of the kernel. Especially the eigenbasis of T in L^2_X does not depend on m. That can be linked with the previous remarks made in Section 4.
2. Assumption (A3) defines here the size of the RKHS: the smaller $\alpha = 2m$ is, the bigger the space is, the harder it is to learn a function.

In the next section, we illustrate on such a toy model our main results and compare our learning algorithm to Ying and Pontil's [47], Tarrès and Yao's [39] and Zhang's [48] algorithms.

5.2. *Experimental set-up.* We use $g(x) = B_{r/2}(x)$ with $\alpha = 2r + 1$, as proposed above, with $B_1(x) = x - \frac{1}{2}$, $B_2(x) = x^2 - x + \frac{1}{6}$ and $B_3(x) = x^3 - \frac{3}{2}x^2 + \frac{1}{2}x$.

We give in Table 2 the functions used for simulations in a few cases that span our three regions. We also remind the choice of γ proposed for the 4 algorithms. We always use the finite horizon setting.

5.3. *Optimal learning rate for our algorithm.* In this section, we empirically search for the best choice of a finite horizon learning rate, in order to check if it matches our prediction. For a certain number of values for n, distributed exponentially between 1 and $10^{3.5}$, we look for the best choice $\gamma_{\mathrm{best}}(n)$ of a constant learning rate for our algorithm up to horizon n. In order to do that, for a large number of constants $C_1, \ldots, C_p$, we estimate the expectation of error $\mathbb{E}[\|\bar{g}_n(\gamma = C_i)) - (g)\|]$ by averaging over 30 independent samples of size n, then report the constant giving minimal error as a function of n in Figure 2. We consider here the situation $\beta = 2, r = 0.75$. We plot results in a logarithmic scale in Figure 2, and evaluate the asymptotic decrease of $\gamma_{\mathrm{best}}(n)$ by fitting an affine approximation to the second half of the curve. We get a slope of -0.51, which matches our choice of -0.5 from Corollary 2. Although, our theoretical results are only upper-bounds, we conjecture that our proof technique also leads to lower-bounds in situations where assumptions (a3)–(a4) hold (like in this experiment).

5.4. *Comparison to competing algorithms.* In this section, we compare the convergence rates of the four algorithms described in Section 4.5. We consider the different choices of (r, α) as described in Table 2 in order to go all over the

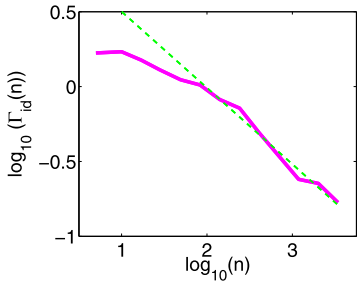


FIG. 2. *Optimal learning rate* $\Gamma_{\text{best}}(n)$ *for our algorithm in the finite horizon setting (plain magenta). The dashed green curve is a first order affine approximation of the second half of the magenta curve.*

different optimality situations. The main properties of each algorithm are described in Table 1. However we may note:

- For our algorithm, $\Gamma(n)$ is chosen accordingly to Corollary 2, with $\gamma_0 = \frac{1}{R^2}$.
- For Ying and Pontil's algorithm, accordingly to Theorem 6 in [47], we consider $\Gamma(n) = \gamma_0 n^{-2r/(2r+1)}$. We choose $\gamma_0 = \frac{1}{R^2}$ which behaves better than the proposed $\frac{r}{64(1+R^4)(2r+1)}$.
- For Tarrès and Yao's algorithm, we refer to Theorem C in [39], and consider $\Gamma(n) = a(n_0+n)^{-2r/(2r+1)}$ and $\Lambda(n) = \frac{1}{a}(n_0+n)^{-1/(2r+1)}$. The theorem is stated for all $a \geq 4$: we choose $a = 4$.
- For Zhang's algorithm, we refer to Part 2.2 in [47], and choose $\Gamma(n) = \gamma_0 n^{-2r/(2r+1)}$ with $\gamma_0 = \frac{1}{R^2}$ which behaves better than the proposed choice $\frac{1}{4(1+R^2)}$.

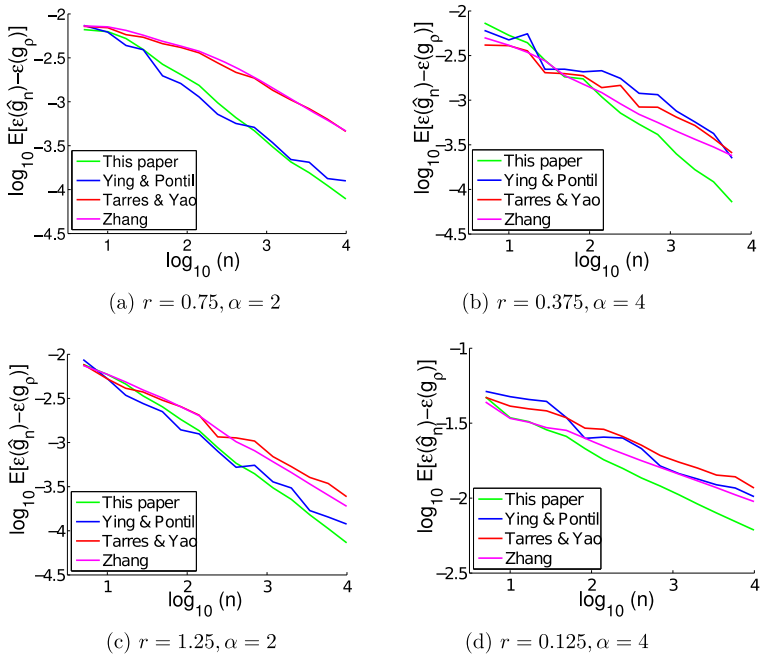
In Figure 3, we plot the expected error as a function of the number of points.

Finally, we sum up the rates that were both predicted and derived for the four algorithms in the four cases for (α, δ) in Table 3. It appears that (a) we approximatively match the predicted rates in most cases (they would if n was larger), (b) our rates improve on existing work.

TABLE 2

Different choices of the parameters α, r *and the corresponding convergence rates and step-sizes. The* (α, δ) *coordinates of the four choices of couple "(kernel, objective function)" are mapped on Figure 1. They are spread over the different optimality regions*

r	α	δ	K	g_ρ	$\frac{\log(\gamma)}{\log(n)}$ **(this paper)**	$\frac{\log(\gamma)}{\log(n)}$ **(previous)**
0.75	2	4	R_1	B_2	$-1/2 = -0.5$	$-3/5 = -0.6$
0.375	4	4	R_2	B_2	0	$-3/7 \simeq -0.43$
1.25	2	6	R_1	B_3	$-3/7 \simeq -0.43$	$-5/7 \simeq -0.71$
0.125	4	2	R_2	B_1	0	$-1/5 = -0.2$



(a) $r = 0.75, \alpha = 2$ (b) $r = 0.375, \alpha = 4$

(c) $r = 1.25, \alpha = 2$ (d) $r = 0.125, \alpha = 4$

FIG. 3. *Comparison between algorithms. We have chosen parameters in each algorithm accordingly to description in Section* 4.5, *especially for the choices of* γ_0. *The y-axis is* $\log_{10}(\mathbb{E}[\varepsilon(\hat{g}_n) - \varepsilon(g_\rho)])$, *where the final output* $\hat{g}_n$ *may be either* $\bar{g}_n$ (*this paper, Zhang*) *or* g_n (*Ying and Pontil, Yao and Tarres*). *This expectation is computed by averaging over* 15 *independent samples*.

6. Conclusion. In this paper, we have provided an analysis of averaged unregularized stochastic gradient methods for kernel-based least-squares regression. Our novel analysis allowed us to consider larger step-sizes, which in turn lead to

TABLE 3

Predicted and effective rates (*asymptotic slope of the* log-log *plot*) *for the four different situations. We leave empty cases when the set-up does not come with existing guarantees: most algorithms seem to exhibit the expected behavior even in such cases*

	$r = 0.75$ $\alpha = 2$	$r = 0.375$ $\alpha = 4$	$r = 1.25$ $\alpha = 2$	$r = 0.125$ $\alpha = 4$
Predicted rate (our algo.)	$\mathbf{-0.75}$	$\mathbf{-0.75}$	$\mathbf{-0.8}$	$\mathbf{-0.25}$
Effective rate (our algo.)	$\mathbf{-0.7}$	$\mathbf{-0.71}$	$\mathbf{-0.69}$	$\mathbf{-0.29}$
Predicted rate (YP)	-0.6	-0.43	-0.71	-0.2
Effective rate (YP)	-0.53	-0.5	-0.63	-0.22
Predicted rate (TY)	-0.6			
Effective rate (TY)	-0.48	-0.39	-0.43	-0.2
Predicted rate (Z)		-0.43		-0.2
Effective rate (Z)	-0.53	-0.43	-0.41	-0.21

optimal estimation rates for many settings of eigenvalue decay of the covariance operators and smoothness of the optimal prediction function. Moreover, we have worked on a more general setting than previous work that includes most interesting cases of positive definite kernels.

Our work can be extended in a number of interesting ways: First, (a) we have considered results in expectation; following the higher-order moment bounds from [5] in the Euclidean case, we could consider higher-order moments, which in turn could lead to high-probability results or almost-sure convergence. Moreover, (b) while we obtain optimal convergence rates for a particular regime of kernels/objective functions, using different types of averaging (i.e., nonuniform) may lead to optimal rates in other regimes. Besides, (c) following [5], we could extend our results for infinite-dimensional least-squares regression to other smooth loss functions, such as for logistic regression, where an online Newton algorithm with the same running-time complexity would also lead to optimal convergence rates. Also, (d) the running-time complexity of our stochastic approximation procedures is still quadratic in the number of samples n, which is unsatisfactory when n is large; by considering reduced set-methods [3, 8, 13], we hope to able to obtain a complexity of $O(d_n n)$, where d_n is such that the convergence rate is $O(d_n/n)$, which would extend the Euclidean space result, where d_n is constant equal to the dimension. Finally, (e) in order to obtain the optimal rates when the bias term dominates our generalization bounds, it would be interesting to combine our spectral analysis with recent accelerated versions of stochastic gradient descent which have been analyzed in the finite-dimensional setting [17].

APPENDIX A: MINIMAL ASSUMPTIONS

A.1. Definitions. We first define the set of square ρ_X-integrable functions $\mathcal{L}^2_{\rho_X}$:

$$\mathcal{L}^2_{\rho_X} = \left\{ f \ : \mathcal{X} \to \mathbb{R} \Big/ \int_{\mathcal{X}} f^2(t)\, d\rho_X(t) < \infty \right\};$$

we will always make the assumptions that this space is separable (this is the case in most interesting situations. See [40] for more details.) $L^2_{\rho_X}$ is its quotient under the equivalence relation given by

$$f \equiv g \quad \Leftrightarrow \quad \int_{\mathcal{X}} \big(f(t) - g(t)\big)^2\, d\rho_X(t) = 0,$$

which makes it a separable Hilbert space (see, e.g., [24]).

We denote p the canonical projection from $\mathcal{L}^2_{\rho_X}$ into $L^2_{\rho_X}$ such that $p : f \mapsto \tilde{f}$, with $\tilde{f} = \{g \in \mathcal{L}^2_{\rho_X}, \text{ s.t. } f \equiv g\}$.

Under assumptions (A1), (A2) or (A1′), (A2′), any function in $\mathcal{H}$ in $\mathcal{L}^2_{\rho_X}$. Moreover, under (A1), (A2) the spaces $\mathcal{H}$ and $p(\mathcal{H})$ may be identified, where $p(\mathcal{H})$ is the image of $\mathcal{H}$ via the mapping $p \circ i : \mathcal{H} \xrightarrow{i} \mathcal{L}^2_{\rho_X} \xrightarrow{p} L^2_{\rho_X}$, where i is the trivial injection from $\mathcal{H}$ into $\mathcal{L}^2_{\rho_X}$.

A.2. Isomorphism. As it has been explained in the main text, the minimization problem will appear to be an approximation problem in $\mathcal{L}^2_{\rho_X}$, for which we will build estimates in $\mathcal{H}$. However, to derive theoretical results, it is easier to consider it as an approximation problem in the Hilbert space $L^2_{\rho_X}$, building estimates in $p(\mathcal{H})$.

We thus need to define a notion of the best estimation in $p(\mathcal{H})$. We first define the closure $\overline{F}$ (with respect to $\|\cdot\|_{L^2_{\rho_X}}$) of any set $F \subset L^2_{\rho_X}$ as the set of limits of sequences in F. The space $\overline{p(\mathcal{H})}$ is a closed and convex subset in $L^2_{\rho_X}$. We can thus define $g_{\mathcal{H}} = \arg\min_{f \in \overline{p(\mathcal{H})}} \varepsilon(g)$, as the orthogonal projection of g_ρ on $\overline{p(\mathcal{H})}$, using the existence of the projection on any closed convex set in a Hilbert space. See Proposition 8 in Appendix A in [15] for details.

PROPOSITION 8 (Definition of best approximation function). *Assume* (A1)–(A2). *The minimum of* $\varepsilon(f)$ *in* $\overline{p(\mathcal{H})}$ *is attained at a certain* $g_{\mathcal{H}}$ (*which is unique and well defined in* $L^2_{\rho_X}$).

Where $\overline{p(\mathcal{H})} = \{f \in L^2_{\rho_X} \setminus \exists (f_n) \subset p(\mathcal{H}), \|f_n - f\|_{L^2_{\rho_X}} \to 0\}$ is the set of functions f for which we can hope for consistency, that is, having a sequence $(f_n)_n$ of estimators in $\mathcal{H}$ such that $\varepsilon(f_n) \to \varepsilon(f)$.

The properties of our estimator, especially its rate of convergence will strongly depend on some properties of both the kernel, the objective function and the distributions, which may be seen through the properties of the covariance operator which is defined in the main text. We have defined the covariance operator, $\Sigma : \mathcal{H} \to \mathcal{H}$. In the following, we extend such an operator as an endomorphism $\mathcal{T}$ from $L^2_{\rho_X}$ to $\mathcal{L}^2_{\rho_X}$ and by projection as an endomorphism $T = p \circ \mathcal{T}$ from $L^2_{\rho_X}$ to $L^2_{\rho_X}$. Note that $\mathcal{T}$ is well defined as $\int_{\mathcal{X}} g(t) K_t \, d\rho_{\mathcal{X}}(t)$ does not depend on the function g chosen in the class of equivalence of g.

DEFINITION 3 (Extended covariance operator). Assume (A1)–(A2). We define the operator $\mathcal{T}$ as follows (this expectation is formally defined as a Bochner expectation in $\mathcal{H}$):

$$\mathcal{T} L^2_{\rho_X} \to \mathcal{L}^2_{\rho_X},$$
$$g \mapsto \int_{\mathcal{X}} g(t) K_t \, d\rho_{\mathcal{X}}(t),$$

so that for any $z \in \mathcal{X}$, $\mathcal{T}(g)(z) = \int_{\mathcal{X}} g(x) K(x,z) \, d\rho_{\mathcal{X}}(t) = \mathbb{E}[g(X)K(X,z)]$.

A first important remark is that $\Sigma f = 0$ implies $\langle f, \Sigma f \rangle = \|f\|^2_{L^2_{\rho_X}} = 0$, that is $p(\mathrm{Ker}(\Sigma)) = \{0\}$. However, Σ may not be injective (unless $\|f\|^2_{L^2_{\rho_X}} \Rightarrow f = 0$, which is true when f is continuous and ρ_X has full support). Σ and $\mathcal{T}$ may independently be injective or not.

The operator T (which is an endomorphism of the separable Hilbert space $L^2_{\rho_X}$) can be reduced in some Hilbertian eigenbasis of $L^2_{\rho_X}$. The linear operator $\mathcal{T}$ happens to have an image included in $\mathcal{H}$, and the eigenbasis of T in $L^2_{\rho_X}$ may also be seen as eigenbasis of Σ in $\mathcal{H}$ (see the proof in Appendix I.2, Proposition 18 in [15]).

PROPOSITION 9 (Decomposition of Σ). *Assume* (A1)–(A2). *The image of* $\mathcal{T}$ *is included in* $\mathcal{H}$: $\mathrm{Im}(\mathcal{T}) \subset \mathcal{H}$, *that is, for any* $f \in L^2_{\rho_X}$, $\mathcal{T}f \in \mathcal{H}$. *Moreover, for any* $i \in I$, $\phi_i^H = \frac{1}{\mu_i}\mathcal{T}\phi_i \in \mathcal{H} \subset \mathcal{L}^2_{\rho_X}$ *is a representant for the equivalence class* ϕ_i, *that is,* $p(\phi_i^H) = \phi_i$. *Moreover* $\mu_i^{1/2}\phi_i^H$ *is an orthonormal eigensystem of the orthogonal supplement* $\mathscr{S}$ *of the null space* $\mathrm{Ker}(\Sigma)$. *That is*:

- $\forall i \in I$, $\Sigma\phi_i^H = \mu_i\phi_i^H$.
- $\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}$.

Such decompositions allow to define $\mathcal{T}^r : L^2_{\rho_X} \to \mathcal{H}$ for $r \geq 1/2$. Indeed, completeness allows to define infinite sums which satisfy a Cauchy criterion. See proof in Appendix I.2, Proposition 19 in [15]. Note the different condition concerning r in the definitions. For $r \geq 1/2$, $T^r = p \circ \mathcal{T}^r$. We need $r \geq 1/2$, because $(\mu_i^{1/2}\phi^H)$ is an orthonormal system of $\mathscr{S}$.

DEFINITION 4 (Powers of $\mathcal{T}$). We define, for any $r \geq 1/2$, $\mathcal{T}^r : L^2_{\rho_X} \to \mathcal{H}$, for any $h \in \mathrm{Ker}(T)$ and $(a_i)_{i\in I}$ such that $\sum_{i\in I} a_i^2 < \infty$, through

$$\mathcal{T}^r\left(h + \sum_{i\in I} a_i\phi_i\right) = \sum_{i\in I} a_i\mu_i^r\phi_i^H.$$

We have two decompositions of $L^2_{\rho_X} = \mathrm{Ker}(T) \overset{\perp}{\oplus} S$ and $\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}$. The two orthogonal supplements S and $\mathscr{S}$ happen to be related through the mapping $\mathcal{T}^{1/2}$, as stated in Proposition 4: $\mathcal{T}^{1/2}$ is an isomorphism from S into $\mathscr{S}$. It also has he following consequences, which generalizes Corollary 1.

COROLLARY 7.

- $T^{1/2}(S) = p(\mathcal{H})$, *that is, any element of* $p(\mathcal{H})$ *may be expressed as* $T^{1/2}g$ *for some* $g \in L^2_{\rho_X}$.
- *For any* $r \geq 1/2$, $T^r(S) \subset \mathcal{H}$, *because* $T^r(S) \subset T^{1/2}(S)$, *that is, with large powers* r, *the image of* T^r *is in the projection of the Hilbert space.*
- $\forall r > 0$, $\overline{T^r(L^2_{\rho_X})} = S = \overline{T^{1/2}(L^2_{\rho_X})} = \overline{\mathcal{H}}$, *because* (a) $T^{1/2}(L^2_{\rho_X}) = p(\mathcal{H})$ *and* (b) *for any* $r > 0$, $\overline{T^r(L^2_{\rho_X})} = S$. *In other words, elements of* $\overline{p(\mathcal{H})}$ (*on which our minimization problem attains its minimum*), *may seen as limits* (*in* $L^2_{\rho_X}$) *of elements of* $T^r(L^2_{\rho_X})$, *for any* $r > 0$.

– $p(\mathcal{H})$ is dense in $L^2_{\rho_X}$ if and only if T is injective.

A.3. Mercer theorem generalized. Finally, although we will not use it in the rest of the paper, we can state a version of Mercer's theorem, which does not make any more assumptions that are required for defining RKHSs.

PROPOSITION 10 (Kernel decomposition). *Assume* (A1)–(A2). *We have for all* $x, y \in \mathcal{X}$,

$$K(x,y) = \sum_{i \in I} \mu_i \phi_i^H(x) \phi_i^H(y) + g(x,y),$$

and we have for all $x \in \mathcal{X}$, $\int_{\mathcal{X}} g(x,y)^2 \, d\rho_X(y) = 0$. *Moreover, the convergence of the series is absolute.*

We thus obtain a version of Mercer's theorem (see Appendix I.5.3 in [15]) without any topological assumptions. Moreover, note that (a) $\mathcal{S}$ is also an RKHS, with kernel $(x,y) \mapsto \sum_{i \in I} \mu_i \phi_i^H(x) \phi_i^H(y)$ and (b) that given the decomposition above, the optimization problem in $\mathcal{S}$ and $\mathcal{H}$ have equivalent solutions. Moreover, considering the algorithm below, the estimators we consider will almost surely build equivalent functions (see Appendix I.4 in [15]). Thus, we could assume without loss of generality that the kernel K is exactly equal to its expansion $\sum_{i \in I} \mu_i \phi_i^H(x) \phi_i^H(y)$.

A.4. Complementary (A6) assumption. Under minimal assumptions, we also have to make a complementary moment assumption:

(A6′) There exists $R > 0$ and $\sigma > 0$ such that $\mathbb{E}[\Xi \otimes \Xi] \preccurlyeq \sigma^2 \Sigma$, and $\mathbb{E}(K(X,X) \times K_X \otimes K_X) \preccurlyeq R^2 \Sigma$ where $\preccurlyeq$ denotes the order between self-adjoint operators.

In other words, for any $f \in \mathcal{H}$, we have: $\mathbb{E}[K(X,X) f(X)^2] \leq R^2 \mathbb{E}[f(X)^2]$. Such an assumption is implied by (A2), that is, if $K(X,X)$ is almost surely bounded by R^2: this constant can then be understood as the radius of the set of our data points. However, our analysis holds in these more general set-ups where only fourth-order moment of $\|K_x\|_{\mathcal{H}} = K(x,x)^{1/2}$ is finite.

APPENDIX B: SKETCH OF THE PROOFS

Our main theorems are Theorem 2 and Theorem 3, respectively, in the finite horizon and in the online setting. Corollaries can be easily derived by optimizing over γ the upper bound given in the theorem.

The complete proof is given in Appendix II in [15]. The proof is nearly the same for finite horizon and online setting. It relies on a refined analysis of strongly related recursions in the RKHS and on a comparison between iterates of the recursions (controlling the deviations).

We first present the sketch of the proof for the *finite-horizon setting*: We want to analyze the error of our sequence of estimators (g_n) such that $g_0 = 0$ and

$$g_n = g_{n-1} - \gamma_n\big[y_n - \langle g_{n-1}, K_{x_n}\rangle_{\mathcal{H}}\big]K_{x_n},$$
$$g_n = (I - \gamma K_{x_n} \otimes K_{x_n})g_{n-1} + \gamma y_n K_{x_n},$$
$$g_n - g_{\mathcal{H}} = (I - \gamma \widetilde{K_{x_n} \otimes K_{x_n}})(g_{n-1} - g_{\mathcal{H}}) + \gamma\,\Xi_n.$$

Where we have denoted $\Xi_n = (y_n - g_{\mathcal{H}}(x_n))K_{x_n}$ the residual, which has 0 mean, and $\widetilde{K_{x_n} \otimes K_{x_n}} : L^2_{\rho_X} \to \mathcal{H}$ an a.s. defined extension of $K_{x_n} \otimes K_{x_n} : \mathcal{H} \to \mathcal{H}$, such that $\widetilde{K_{x_n} \otimes K_{x_n}}(f) = f(x_n)K_{x_n}$, that will be denoted for simplicity $K_{x_n} \otimes K_{x_n}$ in this section.

Finally, we are studying a sequence $(\eta_n)_n = (g_n - g_{\mathcal{H}})_n$ defined by

$$\eta_0 = g_{\mathcal{H}},$$
$$\eta_n = (I - \gamma_n K_{x_n} \otimes K_{x_n})\eta_{n-1} + \gamma_n\,\Xi_n.$$

We first consider splitting this recursion in two simpler recursions η_n^{init} and η_n^{noise} such that $\eta_n = \eta_n^{\text{init}} + \eta_n^{\text{noise}}$:

- $(\eta_n^{\text{init}})_n$ defined by
$$\eta_0^{\text{init}} = g_{\mathcal{H}} \quad \text{and} \quad \eta_n^{\text{init}} = (I - \gamma K_{x_n} \otimes K_{x_n})\eta_{n-1}^{\text{init}}.$$
η_n^{init} is the part of $(\eta_n)_n$ which is due to the *initial conditions* (it is equivalent to assuming $\Xi_n \equiv 0$).
- Respectively, let $(\eta_n^{\text{noise}})_n$ be defined by
$$\eta_0^{\text{noise}} = 0 \quad \text{and} \quad \eta_n^{\text{noise}} = (I - \gamma K_{x_n} \otimes K_{x_n})\eta_{n-1}^{\text{noise}} + \gamma\,\Xi_n.$$
η_n^{noise} is the part of $(\eta_n)_n$ which is due to the *noise*.

We will bound $\|\eta_n\|$ by $\|\eta_n^{\text{init}}\| + \|\eta_n^{\text{noise}}\|$ using Minkowski's inequality. *That is how the bias-variance trade-off originally appears.*

Next, we notice that $\mathbb{E}[K_{x_n} \otimes K_{x_n}] = \mathcal{T}$, and thus define "semi-stochastic" versions of the previous recursions by replacing $K_{x_n} \otimes K_{x_n}$ by its expectation:

For the initial conditions: $(\eta_n^{0,\text{init}})_{n\in\mathbb{N}}$ so that

$$\eta_0^{0,\text{init}} = g_{\mathcal{H}}, \qquad \eta_n^{0,\text{init}} = (I - \gamma\mathcal{T})\eta_{n-1}^{0,\text{init}},$$

which is a deterministic sequence.

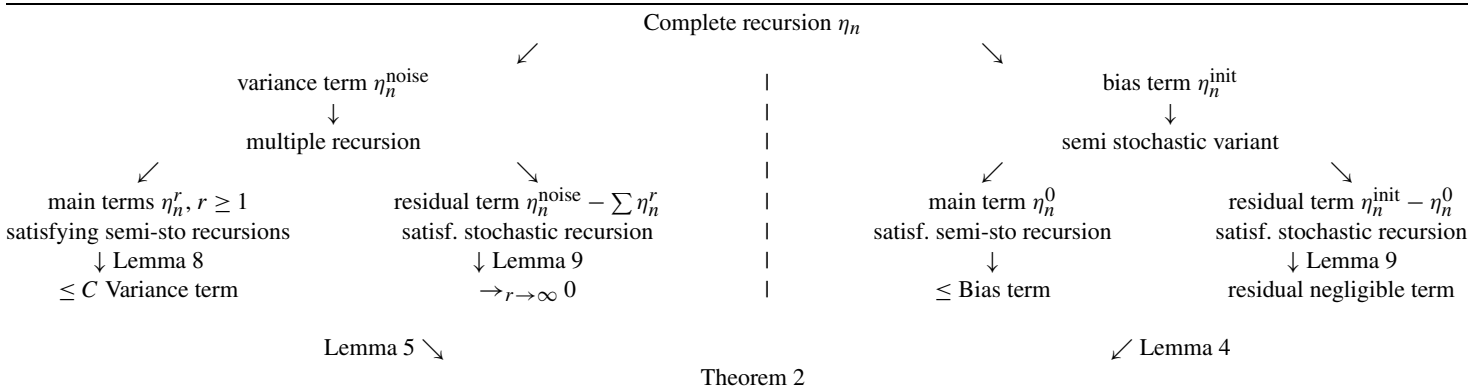
An algebraic calculation gives an estimate of the norm of $\eta_n^{0,\text{init}}$, and we can also bound the residual term $\eta_n^{init} - \eta_n^{0,\text{init}}$, then conclude by Minkowski.

For the variance term: We follow the exact same idea, but have to define a sequence of "semi-stochastic recursion," to be able to bound the residual term.

This decomposition is summed up in Table 4.

For the *online setting*, we follow comparable ideas and end in a similar decomposition.

TABLE 4
Error decomposition in the finite horizon setting. All the references refer to lemmas given in Appendix II *in* [15]



Acknowledgement. We thank Nicolas Flammarion for helpful discussions.

REFERENCES

[1] ABRAMOWITZ, M. and STEGUN, I. (1964). *Handbook of Mathematical Functions*. Dover, New York.

[2] ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404. MR0051437

[3] BACH, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In *Proceedings of the International Conference on Learning Theory* (*COLT*).

[4] BACH, F. and MOULINES, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Adv. NIPS*. Curran Associates, Inc., Granada, Spain.

[5] BACH, F. and MOULINES, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems* (*NIPS*). Curran Associates, Inc., Lake Tahoe, CA.

[6] BERLINET, A. and THOMAS-AGNAN, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston, MA. MR2239907

[7] BLANCHARD, G. and KRÄMER, N. (2010). Optimal learning rates for kernel conjugate gradient regression. In *Advances in Neural Inf. Proc. Systems* (*NIPS*) 226–234. Curran Associates, Inc., Vancouver, Canada.

[8] BORDES, A., ERTEKIN, S., WESTON, J. and BOTTOU, L. (2005). Fast kernel classifiers with online and active learning. *J. Mach. Learn. Res.* **6** 1579–1619. MR2249866

[9] BREZIS, H. (1983). *Analyse Fonctionnelle, Théorie et Applications*. Masson, Paris. MR0697382

[10] CAPONNETTO, A. and DE VITO, E. (2007). Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **7** 331–368. MR2335249

[11] CUCKER, F. and SMALE, S. (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)* **39** 1–49 (electronic). MR1864085

[12] CUCKER, F. and SMALE, S. (2002). Best choices for regularization parameters in learning theory: On the bias-variance problem. *Found. Comput. Math.* **2** 413–428. MR1930945

[13] DEKEL, O., SHALEV-SHWARTZ, S. and SINGER, Y. (2005). The Forgetron: A kernel-based perceptron on a fixed budget. In *Adv. NIPS*. Curran Associates, Inc., Vancouver, Canada.

[14] DE VITO, E., CAPONNETTO, A. and ROSASCO, L. (2005). Model selection for regularized least-squares algorithm in learning theory. *Found. Comput. Math.* **5** 59–85. MR2125691

[15] DIEULEVEUT, A. and BACH, F. (2014). Nonparametric stochastic approximation with large step sizes. Preprint. Available at arXiv:1408.0361.

[16] ENGL, H. W., HANKE, M. and NEUBAUER, A. (1996). *Regularization of Inverse Problems*. *Mathematics and Its Applications* **375**. Kluwer Academic, Dordrecht. MR1408680

[17] FLAMMARION, N. and BACH, F. (2015). From averaging to acceleration, there is only a step-size. In *Proceedings of the International Conference on Learning Theory* (*COLT*). Microtome Publishing, Paris, France.

[18] GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. Johns Hopkins Univ. Press, Baltimore, MD. MR1417720

[19] HAZAN, E. and KALE, S. (2011). Beyond the regret minimization barrier: An optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the International Conference on Learning Theory* (*COLT*). Microtome Publishing, Budapest, Hungary.

[20] HSU, D., KAKADE, S. M. and ZHANG, T. (2014). Random design analysis of ridge regression. *Found. Comput. Math.* **14** 569–600. MR3201956

[21] JOHNSTONE, I. M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. In *Statistical Decision Theory and Related Topics*, *V* (*West Lafayette*, *IN*, 1992) 303–326. Springer, New York. MR1286310

[22] KIMELDORF, G. and WAHBA, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33** 82–95. MR0290013

[23] KIVINEN, J., SMOLA, A. J. and WILLIAMSON, R. C. (2004). Online learning with kernels. *IEEE Trans. Signal Process.* **52** 2165–2176. MR2085578

[24] KOLMOGOROV, A. N. and FOMIN, S. V. (1999). *Elements of the Theory of Functions and Functional Analysis*, *Vol.* 1 Dover, New York.

[25] LACOSTE-JULIEN, S., SCHMIDT, M. and BACH, F. (2012). A simpler approach to obtaining an O(1/t) rate for the stochastic projected subgradient method. Preprint. Available at arXiv:1212.2002.

[26] MAHONEY, M. W. (2011). Randomized algorithms for matrices and data. *Faund. Trends Mach. Learn.* **3** 123–224.

[27] MICCHELLI, C. A., XU, Y. and ZHANG, H. (2006). Universal kernels. *J. Mach. Learn. Res.* **7** 2651–2667. MR2274454

[28] MIKUSINSKI, P. and WEISS, E. (2014). The Bochner integral. Preprint. Available at arXiv:1403.5209.

[29] NEMIROVSKI, A., JUDITSKY, A., LAN, G. and SHAPIRO, A. (2008). Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** 1574–1609. MR2486041

[30] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Early stopping for nonparametric regression: An optimal data-dependent stopping rule. In 49*th Annual Allerton Conference on Communication*, *Control*, *and Computing*. IEEE, Monticello, IL, USA.

[31] ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668

[32] ROSASCO, L., TACCHETTI, A. and VILLA, S. (2014). Regularization by early stopping for online learning algorithms. Preprint. Available at arXiv:1405.0042.

[33] SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.

[34] SHALEV-SHWARTZ, S. (2011). Online learning and online convex optimization. *Foundations and Trends in Machine Learning* **4** 107–194.

[35] SHAWE-TAYLOR, J. and CRISTIANINI, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, Cambridge, MA.

[36] SMALE, S. and ZHOU, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constr. Approx.* **26** 153–172. MR2327597

[37] SRIPERUMBUDUR, B. K., FUKUMIZU, K. and LANCKRIET, G. R. G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.* **12** 2389–2410. MR2825431

[38] STEINWART, I., HUSH, D. and SCOVEL, C. (2009). Optimal rates for regularized least squares regression. In *Proceedings of the* 22*nd Annual Conference in Learning Theory*. Microtome Publishing, Montreal, Canada.

[39] TARRÈS, P. and YAO, Y. (2011). Online learning as stochastic approximation of regularization paths. Preprint. Available at arXiv:1103.5538.

[40] THOMSON, B., BRUCKNER, J. and BRUCKNER, A. M. (2000). *Elementary Real Analysis*. Pearson Education, Upper Saddle River, NJ.

[41] TSYBAKOV, A. B. (2008). *Introduction to Nonparametric Estimation*, 1st ed. Springer, Berlin.

[42] VERT, J. (2014). Kernel methods. Unpublished manuscript.

[43] WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA. MR1045442

[44] WILLIAMS, C. and SEEGER, M. (2001). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems* 13 (T. K. Leen, T. G. Dietterich and V. Tresp, eds.). MIT Press, Cambridge, MA.

[45] YAO, Y. (2006). A dynamic theory of learning. Ph.D. thesis, Univ. of California, Berkeley. MR2709954

[46] YAO, Y., ROSASCO, L. and CAPONNETTO, A. (2007). On early stopping in gradient descent learning. *Constr. Approx.* **26** 289–315. MR2327601

[47] YING, Y. and PONTIL, M. (2008). Online gradient descent learning algorithms. *Found. Comput. Math.* **8** 561–596. MR2443089

[48] ZHANG, T. (2004) Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *ICML* 2014 *Proceedings of the Twenty-First International Conference on Machine Learning*. Microtome Publishing, Beijing, China.

[49] ZHANG, T. and YU, B. (2005). Boosting with early stopping: Convergence and consistency. *Ann. Statist.* **33** 1538–1579. MR2166555

DÉPARTEMENT D'INFORMATIQUE DE L'ECOLE NORMALE SUPÉRIEURE
INRIA / CNRS / ENS
45, RUE D'ULM
75005 PARIS
FRANCE
E-MAIL: aymeric.dieuleveut@ens.fr
francis.bach@inria.fr