SIAM J. NUMER. ANAL.
Vol. 57, No. 5, pp. 2217–2245

ADAPTIVE APPROXIMATION BY OPTIMAL WEIGHTED LEAST-SQUARES METHODS[*]

GIOVANNI MIGLIORATI[†]

Abstract. Given any domain $X \subseteq \mathbb{R}^d$ and a probability measure ρ on X, we study the problem of approximating in $L^2(X, \rho)$ a given function $u : X \to \mathbb{R}$, using its noiseless pointwise evaluations at random samples. For any given linear space $V \subset L^2(X, \rho)$ with dimension n, previous works have shown that stable and optimally converging weighted least-squares (WLS) estimators can be constructed using m random samples distributed according to an auxiliary probability measure μ that depends on V, with m being linearly proportional to n up to a logarithmic term. As a first contribution, we present novel results on the stability and accuracy of WLS estimators with a given approximation space, using random samples that are more structured than those used in the previous analysis. As a second contribution, we study approximation by WLS estimators in the adaptive setting. For any sequence of nested spaces $(V_k)_k \subset L^2(X, \rho)$, we show that a sequence of WLS estimators of u, one for each space V_k, can be sequentially constructed such that (i) the estimators remain provably stable with high probability and optimally converging in expectation, simultaneously for all iterations from one to k, and (ii) the overall number of samples necessary to construct all the first k estimators remains linearly proportional to the dimension of V_k, up to a logarithmic term. The overall number of samples takes into account all the samples generated to build all the estimators from iteration one to k. We propose two sampling algorithms that achieve this goal. The first one is a purely random algorithm that recycles most of the samples from the previous iterations. The second algorithm recycles all the samples from all the previous iterations. Such an achievement is made possible by crucially exploiting the structure of the random samples. Finally we apply the results from our analysis to develop numerical methods for the adaptive approximation of functions in high dimension.

Key words. approximation theory, weighted least squares, convergence rates, high dimensional approximation, adaptive approximation, multivariate polynomials, wavelets

AMS subject classifications. 41A65, 41A25, 41A10, 65T60

DOI. 10.1137/18M1198387

1. Introduction.

In recent years, the increasing computing power and availability of data have contributed to huge growth in the complexity of the mathematical models. Dealing with such models often requires the approximation or integration of functions in high dimension, which can be a challenging task due to the curse of dimensionality. The present paper studies the problem of approximating a function $u : X \to \mathbb{R}$ that depends on a d-dimensional parameter $x \in X \subseteq \mathbb{R}^d$, using the information coming from the evaluations of u at a set of selected samples $x^1, \dots, x^m \in X$. Two classical approaches to such a problem are *interpolation* and *least-squares methods*; see, e.g., [9, 3, 15]. Here we turn our attention to least-squares methods, which are frequently used in applications for approximation, data-fitting, estimation, and prediction. Other approaches to function approximation are *compressive sensing* (see [12] and references therein) and *neural networks* (see, e.g., [4, 16]).

Previous convergence results for standard least-squares methods have been proposed in [5], in expectation, and [17], in probability. Weighted least-squares (WLS) methods have been previously studied in [11, 13, 7]. It has been proven in [7] that stable and optimally converging WLS estimators can be constructed using judiciously

[*] Received by the editors July 6, 2018; accepted for publication (in revised form) July 3, 2019; published electronically September 12, 2019.
https://doi.org/10.1137/18M1198387

[†] Sorbonne Université, Université de Paris, CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, 75005, Paris, France (giovanni.migliorati@gmail.com).

distributed random samples, whose number is only linearly proportional to the dimension of the approximation space, up to a logarithmic term. The analysis holds in general approximation spaces, and the number of samples ensuring stability and optimality of the estimator does not depend on d. Such a result is recalled in Theorem 1. The analysis in [7] considers both cases of noisy or noiseless evaluations of u. This paper is confined to the case of noiseless evaluations, which is relevant whenever the function u can be evaluated at the selected samples with sufficiently high precision, e.g., up to machine epsilon. The case of noisy evaluations can be addressed using the same techniques as in [7, 19].

The proof of Theorem 1, and more generally the analysis in [7], uses results from [1, 20] on tail bounds for sums of random matrices. An interesting feature of the bounds in [20] is that the matrices need not be identically distributed. The analysis in [7] does not take advantage of this property. One of the main goals of the present paper is to show how the use of this property paves the way toward novel results in the analysis of WLS methods for a given space (Theorem 2) and toward their application in an adaptive setting (Theorem 3). The proof of Theorem 2 builds on previous contributions [5, 7]. The overall skeleton of the proof is similar, but with some crucial differences that make use of the additional structure of the random samples.

The outline of the paper is the following. In section 1.1 we describe and motivate our contributions. In section 2 we recall some results from the analysis in [7] on WLS for a given space. Section 2.2 contains Theorem 2 and its proof. In section 3 we apply Theorems 2 and 1 to the adaptive setting, with an arbitrary nested sequence $(V_k)_k$ of approximation spaces. In section 4 we present the sampling algorithms. Section 5 contains some numerical tests, together with an example of adaptive algorithm that uses sequences of nested polynomial spaces. Section 6 draws some conclusions. All the algorithms are collected in the appendix.

1.1. Motivations and outline of the main results.

Let $X \subseteq \mathbb{R}^d$ be a Borel set, ρ be a Borel probability measure on X, $(\psi_i)_{i\geq 1}$ be a basis orthonormal in $L^2(X,\rho)$ equipped with the inner product $\langle f_1, f_2\rangle = \int_X f_1(x)\, f_2(x)\, d\rho$, and $V := \operatorname{span}\{\psi_1,\dots,\psi_n\}$ be the space obtained by retaining n terms of the basis. The least-squares method approximates the function u by computing its discrete L^2 projection onto a given space V, using pointwise evaluations of u at a set of $m \geq n$ distinct random samples $x^1,\dots,x^m$. The analysis in [7], whose main findings are resumed in the forthcoming Theorem 1, provides some results on the stability and convergence properties of such a discrete projection, and of other WLS estimators as well. In Theorem 1, independent and identically distributed (i.i.d.) random samples are drawn from the probability measure

$$d\mu_n = \frac{1}{n}\sum_{j=1}^{n} d\chi_j,$$

which is an additive mixture of the probability measures χ_j defined as

$$\chi_j(A) := \int_A |\psi_j(x)|^2\, d\rho \qquad \text{for any Borel set } A \subseteq X. \tag{1.1}$$

One sample from μ_n can be generated by randomly choosing an index j uniformly in $\{1,\dots,n\}$ and then drawing one sample from χ_j. In general μ_n is not a product measure, even if ρ is a product measure.

Another novel approach proposed in this paper uses a different type of random samples. Such an approach uses a set of independent random samples of the form

$x^1, \dots, x^m$ with $m = \tau n$ for a suitable integer τ and such that for any $j = 1, \dots, n$, the samples $x^{(j-1)\tau+1}, \dots, x^{j\tau}$ are distributed according to χ_j. These samples are not identically distributed. On the upside, they are more structured than those used in Theorem 1, since the amount of samples coming from each component of the mixture is fixed. If $\tau = 1$, then the n independent samples $x^1, \dots, x^n$ are jointly drawn from

$$(x^1, \dots, x^n) \sim d\gamma_n := \prod_{j=1}^{n} |\psi_j(x^j)|^2 \, d\rho.$$

If $\tau \geq 1$ the draw of $m = \tau n$ independent samples $x^1, \dots, x^m$ follows the probability measure $d\gamma^m := \otimes^\tau d\gamma_n$.

Denote with $d\mu^m := \otimes^m d\mu_n$ the probability measure for the draw of m i.i.d. samples from μ_n. Given a fixed n, in the limit $m = \tau n \to \infty$ obtained by $\tau \to \infty$, the proportion of random samples of μ^m coming from each χ_j tends to $1/n$ by the strong law of large numbers, whereas the same proportion is exactly equal to $1/n$ by construction for the samples drawn from γ^m. With any m the two probability measures μ^m and γ^m generate samples with different distributions. However, the block of n samples from γ_n still mimics the samples from μ_n. For example the sum of the expectation of the random samples is preserved, i.e., if $(x^1, \dots, x^n) \sim d\gamma_n$ and $\tilde{x} \sim d\mu_n$ then

$$\sum_{j=1}^{n} \mathbb{E}\left(x^j\right) = \sum_{j=1}^{n} \int_X x^j |\psi_j(x^j)|^2 \, d\rho = \int_X \tilde{x} \sum_{j=1}^{n} |\psi_j(\tilde{x})|^2 \, d\rho = n\mathbb{E}(\tilde{x}),$$

and this preservation plays a main role in our forthcoming analysis. All the measures appearing in this paper are also Borel probability measures, and sometimes for brevity we refer to them just as measures.

The first main result of this paper is Theorem 2. It proves the same guarantees as Theorem 1 for the stability and accuracy of WLS estimators with a given approximation space, but when the random samples from μ^m are replaced with random samples from γ^m. The second main result concerns the analysis of WLS estimators, when considering a sequence of nested approximation spaces $(V_k)_{k\geq 1}$, where $V_k := \mathrm{span}\{\psi_1, \dots, \psi_{n_k}\}$ and $n_k := \dim(V_k)$. In this adaptive setting, we compare the two approaches using random samples from γ^m or μ^m. In both cases, in Theorems 3 and 4, we prove that a sequence of estimators of u, one for each space V_k, can be sequentially constructed such that (i) the estimators remain provably stable with high probability and optimally converging in expectation, simultaneously for all iterations from one to k, and (ii) the overall number of samples necessary to construct all the first k estimators remains linearly proportional to the dimension of V_k, up to a logarithmic term. As a further contribution we show that using the samples from γ^m rather than from μ^m provides the following advantages, which are relevant in the development of adaptive WLS methods.

- **Structure of the random samples**. When using γ^m, the number of random samples coming from each component $|\psi_j(x)|^2 \, d\rho$ of the mixture is precisely determined and allows the development of adaptive algorithms that recycle all the samples from all the previous iterations. When using μ^m it is not possible to recycle all the samples from the previous iterations with probability equal to one. Given two nested spaces $V_{k-1} \subset V_k$ and two positive integers $\tau_{k-1} \leqslant \tau_k$, at iteration k the probability measure γ^{m_k} of the $m_k = \tau_k n_k$ samples can be decomposed as

(1.2)

$$
\begin{aligned}
d\gamma^{m_k} &= \otimes^{\tau_k} \prod_{j=1}^{n_k} |\psi_j(x)|^2 \, d\rho \\
&= \left(\underbrace{d\gamma^{m_{k-1}}}_{\substack{\text{measure of the } m_{k-1} \\ \text{samples recycled with} \\ \text{certainty from step } k-1}} \right) \otimes \left(\underbrace{\otimes^{\tau_k - \tau_{k-1}} \prod_{j=1}^{n_{k-1}} |\psi_j(x)|^2 \, d\rho}_{\substack{\text{measure of the new samples drawn} \\ \text{from the old components} \\ \text{of the mixture}}} \right) \otimes \left(\underbrace{\otimes^{\tau_k} \prod_{j=1+n_{k-1}}^{n_k} |\psi_j(x)|^2 \, d\rho}_{\substack{\text{measure of the new samples drawn} \\ \text{from the new components} \\ \text{of the mixture}}} \right).
\end{aligned}
$$

The probability measure $d\mu^{m_k} = \otimes^{m_k} d\mu_{n_k}$ cannot be decomposed as the product of two probability measures with one being $\mu^{m_{k-1}}$, because μ_{n_k} is not a product measure. It is, however, possible to leverage the structure of μ_{n_k} as an additive mixture of $\mu_{n_{k-1}}$ and a suitable probability measure σ_{n_k} and to decompose μ^{m_k} as

(1.3)

$$
d\mu^{m_k} = \otimes^{m_k} \left(\frac{n_{k-1}}{n_k} \underbrace{\overbrace{\frac{1}{n_{k-1}} \sum_{j=1}^{n_{k-1}} |\psi_j(x)|^2 \, d\rho}^{d\mu_{n_{k-1}}}}_{\substack{\text{measure of the samples drawn} \\ \text{from the old components of } d\mu_{n_k}, \\ \text{perhaps recycled from step } k-1}} + \frac{n_k - n_{k-1}}{n_k} \underbrace{\overbrace{\frac{1}{n_k - n_{k-1}} \sum_{j=1+n_{k-1}}^{n_k} |\psi_j(x)|^2 \, d\rho}^{d\sigma_{n_k}}}_{\substack{\text{measure of the samples drawn} \\ \text{from the new components of } d\mu_{n_k}}} \right).
$$

When drawing m_k samples from μ_{n_k}, the amount of samples coming from one of the components of $\mu_{n_{k-1}}$ is a binomial random variable with number of trials m_k and rate of success n_{k-1}/n_k for each trial. Whenever this random variable takes values less than m_{k-1}, which always occurs with some positive probability, it is not possible to recycle all the m_{k-1} samples from iteration $k-1$.

- **Variance reduction.** Random mixture proportions induce extra variance in the generated samples. As a consequence, random samples from γ^m are more disciplined than random samples from μ^m. This stabilization effect amplifies when using basis elements whose supports are more localized than globally supported orthogonal polynomials. We say more on this in Remark 6.
- **Coarsening and extension to nonnested sequences of approximation spaces.** When using the samples from γ^m, thanks to the decomposition (1.2), it is possible to remove an element of the basis ψ_j from the space V as well as its associated samples $x^{(j-1)\tau+1}, \dots, x^{j\tau}$ from the whole set $x^1, \dots, x^m$ of $m = \tau n$ samples and at the same time recycle all the $\tau(n-1)$ remaining samples for $V \setminus \{\psi_j\}$. More generally, the use of γ^m allows the development of efficient adaptive methods with arbitrary sequences of approximation spaces $(V_k)_k$ that probe any $\psi_j \notin V_k$ chosen according to some criterion. The method then either retains ψ_j as $V_{k+1} = V_k \cup \{\psi_j\}$ or discards it depending on its contribution to the reduction of (an estimator of) the error from V_k to V_{k+1}.

Comparison with [2]. Section 4.2 presents an analysis of a sampling algorithm (Algorithm 2) that sequentially generates m random samples from μ^m with an arbitrary nested sequence of approximation spaces $(V_k)_k$. In [2] a similar algorithm that uses μ^m has been proposed and analyzed.

Notation for product of measures. Let ρ_1, ρ_2 be two Borel measures on $X \subseteq \mathbb{R}^d$ with the Borel σ-algebra $\mathfrak{B} = \mathfrak{B}(X)$. The notation $\rho_1 \otimes \rho_2$ denotes the product measure on $X \times X$ with the tensor product Borel σ-algebra $\mathfrak{B} \otimes \mathfrak{B}$, that satisfies

$$\rho_1 \otimes \rho_2(A_1 \times A_2) = \rho_1(A_1)\rho_2(A_2) \quad \text{for any } A_1, A_2 \in \mathfrak{B}.$$

2. Optimal weighted least squares for a given approximation space.

2.1. Previous results. Let $X \subseteq \mathbb{R}^d$ be a Borel set and ρ be a Borel probability measure on X. We define the $L^2(X, \rho)$ inner product

$$\langle f_1, f_2 \rangle = \int_X f_1(x)\, f_2(x)\, d\rho(x) \tag{2.1}$$

associated with the norm $\|f\| := \langle f, f \rangle^{1/2}$. Throughout the paper we denote by $(\psi_i)_{i \geq 1}$ an $L^2(X, \rho)$ orthonormal basis. We define the linear space $V :=$ span $\{\psi_1, \ldots, \psi_n\}$ that contains n arbitrarily chosen elements of the basis and denote with $n := \dim(V)$ its dimension. We further assume that

$$\text{for any } x \in X \text{ there exists } \psi_j \in V \text{ such that } \psi_j(x) \neq 0. \tag{2.2}$$

This assumption is verified, for example, if the space V contains the functions that are constant over X. For any given V, we define the weight function $w : X \to \mathbb{R}$ as

$$w(x) := \frac{n}{\sum_{i=1}^n |\psi_i(x)|^2}, \quad x \in X, \tag{2.3}$$

whose denominator does not vanish under assumption (2.2). The function w is known as the Christoffel function, up to a renormalization, when V is the space of algebraic polynomials with prescribed total degree. Using w we define the probability measure

$$d\mu_n := w^{-1} d\rho = \frac{\sum_{i=1}^n |\psi_i(x)|^2}{n} d\rho, \tag{2.4}$$

which depends on the chosen approximation space V. Another inner product used in this paper is

$$\langle f_1, f_2 \rangle_m := \frac{1}{m} \sum_{j=1}^m w(x^j) f_1(x^j) f_2(x^j), \tag{2.5}$$

where the functions w, f_1, f_2 are evaluated at m samples $x^1, \ldots, x^m$ i.i.d. as μ_n. This inner product is associated with the discrete seminorm $\|f\|_m := \langle f, f \rangle_m^{1/2}$. The discrete inner product (2.5) mimics (2.1) due to (2.4). The exact L^2 projection on V of any function $u \in L^2(X, \rho)$ is defined as

$$\Pi_n u := \underset{v \in V}{\operatorname{argmin}} \|u - v\|.$$

In practice such a projection cannot be computed out of very particular cases, motivating the interest toward the discrete least-squares approach. We define the WLS estimator u_W of u as

$$u_W := \Pi_n^m u = \underset{v \in V}{\operatorname{argmin}} \|u - v\|_m,$$

which is obtained by applying the discrete projector Π_n^m on V to u. The estimator u_W is associated to the solution of the linear system

$$Ga = h, \tag{2.6}$$

where the Gramian matrix G and the right-hand side h are defined elementwise as

$$G_{ij} = \langle \psi_i, \psi_j \rangle_m, \quad h_i = \langle u, \psi_i \rangle_m, \quad i, j = 1, \dots, n,$$

and $a = (a_1, \dots, a_n)^\top$ is the vector containing the coefficients of $u_W = \sum_{i=1}^n a_i \psi_i$ expanded over the orthonormal basis. The linear system (2.6) always has at least one solution, which is unique when G is nonsingular. When G is singular we can define u_W as the estimator associated to the unique minimal ℓ_2-norm solution to (2.6). Moreover, we define the $L^2(X, \rho)$ best approximation of u in V as

$$e_{n,2}(u) := \min_{v \in V} \|u - v\| = \|u - \Pi_n u\| \tag{2.7}$$

and the weighted $L^\infty(X, \rho)$ best approximation of u as

$$e_{n,\infty,w}(u) := \inf_{v \in V} \sup_{y \in X} \sqrt{w(y)} |u(y) - v(y)|.$$

Notice that Π_n, Π_n^m, $e_{n,2}$, and $e_{n,\infty,w}$ depend on the chosen space V, not only on its dimension n. The identity matrix is denoted with $I \in \mathbb{R}^{n \times n}$. The spectral norm of any matrix $A \in \mathbb{R}^{n \times n}$ is defined as

$$|||A||| := \sup_{\|v\|_{\ell_2} = 1} \|Av\|_{\ell_2},$$

using the Euclidean inner product in $\mathbb{R}^n$ and its associated norm. Another WLS estimator introduced in [7] is the conditioned estimator:

$$u_C := \begin{cases} u_W & \text{if } |||G - I||| \le \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

One of the main results from [7] is the following theorem; see [7, Theorem 2.1 and Corollary 2.2].

THEOREM 1. *For any real $r > 0$, if the integers m and n are such that the condition*

$$n\,\theta_r \le \frac{m}{\ln m}, \quad \textit{with} \quad \theta_r := \theta^{-1}(1 + r), \quad \theta := \frac{3\ln(3/2) - 1}{2} \approx 0.108, \tag{2.9}$$

is fulfilled, and $x^1, \dots, x^m$ are i.i.d. random samples from μ_n, then the following holds:

(i) *the matrix G satisfies the tail bound*

$$\Pr\left\{ |||G - I||| > \frac{1}{2} \right\} \le 2nm^{-(r+1)} \le 2m^{-r};$$

(ii) *if $u \in L^2(X, \rho)$, then the estimator u_C satisfies*

$$\mathbb{E}(\|u - u_C\|^2) \le (1 + \varepsilon(m)) e_{n,2}(u)^2 + 2\|u\|^2 m^{-r},$$

where $\varepsilon(m) := \frac{4}{\theta_r \ln(m)} \to 0$ as $m \to +\infty$, and θ_r as in (2.9);

(iii) *with probability larger than* $1-2m^{-r}$, *the estimator* u_W *satisfies*

$$\|u-u_W\| \le (1+\sqrt{2})e_{n,\infty,w}(u) \tag{2.10}$$

for all u *such that* $\|\sqrt{w}u\|_{L^\infty} < +\infty$.

The above theorem can be rewritten for a chosen confidence level by setting $\alpha = 2nm^{-(r+1)}$ and replacing the corresponding r in (2.9). For convenience we rewrite condition (2.9) with equality using the ceiling operator, since the number of samples is an integer and usually one wishes to minimize the number of samples m satisfying (2.9) for a given n.

Corollary 1. *For any* $\alpha \in (0,1)$ *and any integer* $n \ge 1$, *if*

$$m = \left\lceil \frac{n}{\theta} \ln\left(\frac{2n}{\alpha}\right)\right\rceil, \quad \text{with } \theta \text{ as in (2.9)}, \tag{2.11}$$

and $x^1, \dots, x^m$ *are* m *i.i.d. random samples from* μ_n, *then*

$$Pr\left(|||G-I||| > \frac{1}{2}\right) \le \alpha.$$

When the evaluations of the function u are noiseless, convergence estimates in probability with confidence level $1-\alpha$ are immediate to obtain. If the evaluations of u are noisy, then convergence estimates in probability of the form (2.10) can still be obtained by using techniques from large deviations to estimate the additional terms due to the presence of the noise, as shown in [19] for standard least squares.

2.2. Novel results. The proof of Theorem 1, and more generally the analysis in [5, 7], uses a result from [1, 20] on tail bounds for sums of random matrices. We recall below this result from [20, Theorem 1.1], in a less general form that simplifies the presentation and still fits our purposes. If $X^1, \dots, X^m$ are independent $n \times n$ random self-adjoint and positive semidefinite matrices satisfying $\lambda_{\max}(X^j) = |||X^j||| \le R$ almost surely and $\mathbb{E}(\sum_{j=1}^m X^j) = I$, then it holds that

$$\Pr\left(\lambda_{\min}\left(\sum_{j=1}^m X^j\right) \le 1-\delta\right) \le n\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\frac{1}{R}}, \quad \delta \in [0,1], \tag{2.12}$$

$$\Pr\left(\lambda_{\max}\left(\sum_{j=1}^m X^j\right) \ge 1+\delta\right) \le n\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\frac{1}{R}}, \quad \delta \ge 0. \tag{2.13}$$

Since for $\delta \in (0,1)$ the upper bound in (2.13) is always greater than or equal to the upper bound in (2.12), it holds that

$$\Pr\left(\left|\left|\left|\sum_{j=1}^m X^j - I\right|\right|\right| > \delta\right) \le 2n\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\frac{1}{R}}. \tag{2.14}$$

Finding a suitable value for R and taking $\delta = \frac{1}{2}$ leads to item (i) in Theorem 1; see [7] for the proof.

One of the features of the bounds (2.12)–(2.13) is that the matrices $X^1, \dots, X^m$ need not be identically distributed. This property has not been exploited in the analysis in [7]. The first contribution of this paper is the following Theorem 2, which states a similar result as Theorem 1, but using a different type of random samples that is very advantageous in itself as well as for the forthcoming application to the adaptive setting.

THEOREM 2. *For any $\alpha \in (0,1)$ and any integer $n \geq 1$, if*

$$m = \tau n, \qquad \text{with } \tau := \left\lceil \theta^{-1} \ln\left(\frac{2n}{\alpha}\right)\right\rceil, \quad \theta \text{ as in (2.9)}, \tag{2.15}$$

and $x^1, \dots, x^{n\tau}$ is a set of independent random samples such that for any $j = 1, \dots, n$ the samples $x^{(j-1)\tau+1}, \dots, x^{j\tau}$ are identically distributed according to χ_j defined in (1.1) *then the following holds:*

(i) *the matrix G satisfies the tail bound*

$$\Pr\left\{|||G - I||| > \frac{1}{2}\right\} \leq \alpha; \tag{2.16}$$

(ii) *if $u \in L^2(X,\rho)$, then the estimator u_C satisfies*

$$\mathbb{E}(\|u - u_C\|^2) \leq \left(1 + \frac{4\theta}{\ln(2n/\alpha)}\right) e_{n,2}(u)^2 + \alpha\|u\|^2; \tag{2.17}$$

(iii) *with probability larger than $1-\alpha$, the estimator u_W satisfies*

$$\|u - u_W\| \leq (1+\sqrt{2})e_{n,\infty,w}(u)$$

for all u such that $\|\sqrt{w}u\|_{L^\infty} < +\infty$.

Proof. Proof of (i). The matrix G can be decomposed as $G = \sum_{j=1}^{n}\sum_{k=1}^{\tau} X^{jk}$, where the X^{jk}, $j = 1, \dots, n$, $k = 1, \dots, \tau$, are mutually independent and, given any $j = 1, \dots, n$, the $X^{j1}, \dots, X^{j\tau}$ are identically distributed copies of the rank-one random matrix $X(x)$ defined elementwise as

$$X_{pq}(x) = \frac{1}{\tau n} w(x)\psi_p(x)\psi_q(x), \quad p, q = 1, \dots, n,$$

with x being a random variable distributed according to χ_j. Notice that the X^{jk}, $j = 1, \dots, n$, $k = 1, \dots, \tau$, are not identically distributed. Anyway, using (2.3), it holds that for any $p, q = 1, \dots, n$,

$$\begin{aligned}
\mathbb{E}(G_{pq}) &= \mathbb{E}\left(\sum_{j=1}^{n}\sum_{k=1}^{\tau} X_{pq}^{jk}\right) \\
&= \sum_{k=1}^{\tau}\sum_{j=1}^{n} \mathbb{E}\left(X_{pq}^{jk}\right) \\
&= \sum_{k=1}^{\tau}\sum_{j=1}^{n} \int_X \frac{1}{\tau n} w(x)\psi_p(x)\psi_q(x)\psi_j(x)^2 d\rho \\
&= \frac{1}{n}\int_X w(x)\psi_p(x)\psi_q(x)\sum_{j=1}^{n}\psi_j(x)^2 d\rho \\
&= \int_X \psi_p(x)\psi_q(x) d\rho \\
&= \delta_{pq},
\end{aligned}$$

and therefore $\mathbb{E}(G) = I$. We then use (2.14) to obtain that if $|||X^{jk}(x)||| \leq R$ almost surely for any $j = 1, \dots, n$ and any $k = 1, \dots, \tau$, then for any $\delta \in (0, 1)$ it holds that

$$\Pr\left(|||G - I||| > \delta\right) \leq 2n \exp\left(-\frac{c_\delta}{R}\right)$$

with $c_\delta := (1 + \delta)\ln(1 + \delta) - \delta > 0$. We choose $\delta = \frac{1}{2}$ and obtain $c_{\frac{1}{2}} = \theta$ as in (2.9). Since X^{jk} has rank one and

$$|||X^{jk}(x)|||^2 = \operatorname{trace}\left((X^{jk}(x))^\top X^{jk}(x)\right) = \left(\frac{1}{\tau n} w(x) \sum_{\ell=1}^{n} \psi_\ell(x)^2\right)^2 = \frac{1}{\tau^2}$$

for all $j = 1, \dots, n$, for all $k = 1, \dots, \tau$, and uniformly for all $x \in X$, we can take $R = 1/\tau$ and obtain that if m and n satisfy (2.15), then

$$\Pr\left(|||G - I||| > \frac{1}{2}\right) \leq 2ne^{-\theta\tau} \leq 2ne^{-\ln(2n/\alpha)} = \alpha.$$

The overall structure of the proof of (ii) follows [7], with some differences due to the fact that here the samples $x^1, \dots, x^m$ are not identically distributed. First we identify the underlying probability measure associated to these samples. The $m = \tau n$ samples $x^1, \dots, x^m$ are all mutually independent and are subdivided into τ blocks, where each block contains n random samples. More precisely, each block contains one random sample distributed as χ_j, for $j = 1, \dots, n$. The probability measure of each block $z = (z^1, \dots, z^n)$ is $d\gamma_n := \prod_{j=1}^{n} |\psi_j(z^j)|^2 \, d\rho$, where each $z^j \in X$. The probability measure of τ blocks, with all the τn random samples $x^1, \dots, x^m$, is $d\gamma^m := \otimes^\tau d\gamma_n$. Let Ω be the set of all possible draws from γ^m, Ω_+ be the set of all draws such that

$$|||G - I||| \leq \frac{1}{2},$$

and $\Omega_- := \Omega \setminus \Omega_+$ be its complement. Under the assumptions of Theorem 2, from (2.16) it holds that

$$\Pr(\Omega_-) = \int_{\Omega_-} d\gamma^m \leq \alpha.$$

Denote $g := u - \Pi_n u$. We consider the event $|||G - I||| \leq \frac{1}{2}$, where it holds that

$$\|u - u_C\|^2 = \|u - u_W\|^2 = \|g\|^2 + \|\Pi_n^m g\|^2$$

since $\Pi_n^m \Pi_n u = \Pi_n u$ and g is orthogonal to V. Denoting with $(a_1, \dots, a_n)^\top$ the solution to the linear system $Ga = b$, and $b = (\langle g, \psi_k \rangle_m)_{k=1,\dots,n}$, we have that

$$\|u - u_C\|^2 = e_{n,2}(u)^2 + \sum_{k=1}^{n} |a_k|^2.$$

Since $|||G - I||| \leq \frac{1}{2} \implies |||G||| \geq \frac{1}{2} \implies |||G^{-1}||| \leq 2$, from the line above

$$\|u - u_C\|^2 \leq e_{n,2}(u)^2 + 4 \sum_{k=1}^{n} |\langle g, \psi_k \rangle_m|^2.$$

In the event $|||G-I|||>\frac{1}{2}$ by the definition of u_C in (2.8) we have $\|u-u_C\|=\|u\|$. Taking the expectation of $\|u-u_C\|^2$ w.r.t. γ^m we obtain that

$$\begin{aligned}\mathbb{E}(\|u-u_C\|^2) &= \int_{\Omega_+}\|u-u_C\|^2\,d\gamma^m+\int_{\Omega_-}\|u-u_C\|^2\,d\gamma^m\\ &\leq \left(e_{n,2}(u)^2+4\sum_{k=1}^{n}\mathbb{E}(|\langle g,\psi_k\rangle_m|^2)\right)\Pr(\Omega_+)+\|u\|^2\Pr(\Omega_-)\\ &\leq e_{n,2}(u)^2+4\sum_{k=1}^{n}\mathbb{E}(|\langle g,\psi_k\rangle_m|^2)+\alpha\|u\|^2.\end{aligned}$$

We now study the second term in the above expression, crucially exploiting the structure of the random samples and the fact that their expectations still pile up and simplify, although the samples are not identically distributed:

$$\begin{aligned}\sum_{k=1}^{n}\mathbb{E}(|\langle g,\psi_k\rangle_m|^2) =&\sum_{k=1}^{n}\mathbb{E}\left(\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}w(x^i)w(x^j)g(x^i)g(x^j)\psi_k(x^i)\psi_k(x^j)\right)\\ =&\frac{1}{m^2}\sum_{k=1}^{n}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbb{E}\left(w(x^i)w(x^j)g(x^i)g(x^j)\psi_k(x^i)\psi_k(x^j)\right)\\ =&\frac{1}{m^2}\left(\underbrace{\sum_{k=1}^{n}\sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq i}}^{m}\mathbb{E}\left(w(x^i)w(x^j)g(x^i)g(x^j)\psi_k(x^i)\psi_k(x^j)\right)}_{I}\right.\\ &\left.+\underbrace{\sum_{k=1}^{n}\sum_{i=1}^{m}\mathbb{E}\left(\left(w(x^i)g(x^i)\psi_k(x^i)\right)^2\right)}_{II}\right).\end{aligned}$$

For term I: with any $k=1,\dots,n$, using in sequence the independence of the samples, the structure of the samples, and the definition of w we obtain

$$\begin{aligned}I=&\sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq i}}^{m}\mathbb{E}\left(w(x^i)g(x^i)\psi_k(x^i)\right)\mathbb{E}\left(w(x^j)g(x^j)\psi_k(x^j)\right)\\ =&\sum_{i=1}^{m}\mathbb{E}\left(w(x^i)g(x^i)\psi_k(x^i)\right)\sum_{\substack{j=1\\ j\neq i}}^{m}\mathbb{E}\left(w(x^j)g(x^j)\psi_k(x^j)\right)\\ =&\sum_{i=1}^{m}\mathbb{E}\left(w(x^i)g(x^i)\psi_k(x^i)\right)\left(\sum_{j=1}^{m}\mathbb{E}\left(w(x^j)g(x^j)\psi_k(x^j)\right)-\mathbb{E}\left(w(x^i)g(x^i)\psi_k(x^i)\right)\right)\\ =&\sum_{i=1}^{m}\mathbb{E}\left(w(x^i)g(x^i)\psi_k(x^i)\right)\left(\sum_{\ell=1}^{\tau}\sum_{j=1}^{n}\int_X w(x)g(x)\psi_k(x)\psi_j(x)^2d\rho-\mathbb{E}\left(w(x^i)g(x^i)\psi_k(x^i)\right)\right)\\ =&\sum_{i=1}^{m}\mathbb{E}\left(w(x^i)g(x^i)\psi_k(x^i)\right)\left(\sum_{\ell=1}^{\tau}\int_X w(x)g(x)\psi_k(x)\sum_{j=1}^{n}\psi_j(x)^2d\rho-\mathbb{E}\left(w(x^i)g(x^i)\psi_k(x^i)\right)\right)\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{m} \mathbb{E}\left(w(x^i)g(x^i)\psi_k(x^i)\right)\left(\tau n \underbrace{\int_X g(x)\psi_k(x)d\rho}_{=0} - \mathbb{E}\left(w(x^i)g(x^i)\psi_k(x^i)\right)\right)\\
&= -\sum_{i=1}^{m}\left(\mathbb{E}\left(w(x^i)g(x^i)\psi_k(x^i)\right)\right)^2 < 0,
\end{aligned}
$$

where $\langle g, \psi_k\rangle = 0$ for all $k = 1, \dots, n$ because g is orthogonal to V. For term II, again exploiting the structure of the samples and the definition of w it holds that

$$
\begin{aligned}
II &= \sum_{i=1}^{m} \mathbb{E}\left(w(x^i)^2 g(x^i)^2 \sum_{k=1}^{n} \psi_k(x^i)^2\right)\\
&= n\sum_{i=1}^{m} \mathbb{E}\left(w(x^i)g(x^i)^2\right)\\
&= n\sum_{j=1}^{\tau}\sum_{k=1}^{n}\int_X w(x)g(x)^2\psi_k(x)^2 d\rho\\
&= n\sum_{j=1}^{\tau}\int_X w(x)g(x)^2\sum_{k=1}^{n}\psi_k(x)^2 d\rho\\
&= n^2\tau\int_X g(x)^2 d\rho\\
&= n^2\tau\|g\|^2.
\end{aligned}
$$

Putting the pieces together, replacing $m = \tau n$ in term II and neglecting the nonpositive contribution of term I, we obtain

$$
\mathbb{E}(\|u - u_C\|^2) \leq \left(1 + \frac{4n}{m}\right) e_{n,2}(u)^2 + \alpha\|u\|^2.
$$

Since $n/m = \tau^{-1} \leq \theta/\ln(2n/\alpha)$ we finally obtain (2.17).

The proof of (iii) uses (i) and then proceeds in the same way as for the proof of (iii) in Theorem 1 from [7]. From the definition of the spectral norm

$$
|||G - I||| \leq \frac{1}{2} \iff \frac{1}{2}\|v\|^2 \leq \|v\|_m^2 \leq \frac{3}{2}\|v\|^2, \quad v \in V,
$$

and this norm equivalence holds at least with probability $1 - \alpha$ from item (i) under condition (2.15). Using the above norm equivalence, the Pythagorean identity $\|u - v\|_m^2 = \|v - u_W\|_m^2 + \|u - u_W\|_m^2$, and $\max(\|u - v\|_m, \|u - v\|) \leq \|\sqrt{w}(u - v)\|_{L^\infty}$, for any $v \in V$ it holds that

$$
\begin{aligned}
\|u - u_W\| &\leq \|u - v\| + \|v - u_W\|\\
&\leq \|u - v\| + \sqrt{2}\|v - u_W\|_m\\
&\leq \|u - v\| + \sqrt{2}\|u - v\|_m\\
&\leq (1 + \sqrt{2})\|\sqrt{w}(u - v)\|_{L^\infty}.
\end{aligned}
$$

Since v is arbitrary we obtain the thesis. ◻

The next corollary, Corollary 2, extends Theorem 2 to any m (not necessarily an integer multiple of n) satisfying (2.11). For any (fixed) $\tau = 0, \dots \lfloor m/n \rfloor$, the set of m random samples in Corollary 2 is obtained by merging $m - \tau n$ random samples distributed as μ_n and τ random samples from χ_j for all $j = 1, \dots, n$. When m is an integer multiple of n and $\tau = \lfloor m/n \rfloor$, Corollary 2 gives Theorem 2 as a particular case. When $\tau = 0$, all the random samples in Corollary 2 are distributed as μ_n, like in Corollary 1.

COROLLARY 2. *For any $\alpha \in (0,1)$, any integers $n \geq 1$, $m \geq n$, and $\tau = 0, \dots, \lfloor m/n \rfloor$, if m satisfies* (2.11) *and $x^1, \dots, x^m$ is a set of independent random samples such that for any $j = 1, \dots, n$ the $x^{(j-1)\tau+1}, \dots, x^{j\tau}$ are identically distributed according to χ_j defined in* (1.1), *and the $x^{n\tau+1}, \dots, x^m$ are identically distributed as μ_n, then items* (i), (ii), *and* (iii) *of Theorem* 2 *hold true.*

Proof. We proceed as in the proof of Theorem 2 but using the decomposition $G = \sum_{j=1}^{n} \sum_{k=1}^{\tau} X^{jk} + \sum_{\ell=n\tau+1}^{m} X^{\ell}$, where all the X^{jk} and X^{ℓ} are mutually independent, and given any $j = 1, \dots, n$, the $X^{j1}, \dots, X^{j\tau}$ are identically distributed copies of the rank-one random matrix $X(x)$ defined elementwise as

$$X_{pq}(x) = \frac{1}{m} w(x) \psi_p(x) \psi_q(x), \quad p, q = 1, \dots, n,$$

with x being a random variable distributed according to χ_j, and the X^{ℓ} are identically distributed copies of $X(x)$ but with x being a random variable distributed according to μ_n. For any $p, q = 1, \dots, n$, $\tau = 0, \dots, \lfloor m/n \rfloor$,

$$\begin{aligned}
\mathbb{E}(G_{pq}) &= \sum_{k=1}^{\tau} \sum_{j=1}^{n} \mathbb{E}\left(X_{pq}^{jk}\right) + \sum_{\ell=n\tau+1}^{m} \mathbb{E}\left(X_{pq}^{\ell}\right) \\
&= \sum_{k=1}^{\tau} \sum_{j=1}^{n} \int_X \frac{1}{m} w(x) \psi_p(x) \psi_q(x) \psi_j(x)^2 d\rho \\
&\quad + \sum_{\ell=n\tau+1}^{m} \int_X \frac{1}{m} w(x) \psi_p(x) \psi_q(x) \frac{\sum_{j=1}^{n} \psi_j(x)^2}{n} d\rho \\
&= \frac{n\tau}{m} \int_X \psi_p(x) \psi_q(x) d\rho + \frac{m - n\tau}{m} \int_X \psi_p(x) \psi_q(x) d\rho = \delta_{pq}.
\end{aligned}$$

When $\tau = 0$ (m is an integer multiple of n and $\tau = \lfloor m/n \rfloor$), the leftmost (rightmost) sum in the first equation above is empty. Since

$$|||X^{jk}(x)||| = |||X^{\ell}(x)||| = \frac{n}{m}$$

for all $j = 1, \dots, n$, for all $k = 1, \dots, \tau$, for all $\ell = n\tau + 1, \dots, m$ and uniformly for all $x \in X$, we can take $R = n/m$ and obtain that if m and n satisfy (2.11), then (2.16) holds true.

For the proof of (2.17), we proceed as in the proof of Theorem 2 with two differences. When bounding term I, we split the random samples as

$$\begin{aligned}\sum_{j=1}^{m}\mathbb{E}(w(x^j)g(x^j)\psi_k(x^j)) &= \sum_{\ell=1}^{\tau}\sum_{j=1}^{n}\int_X w(x)g(x)\psi_k(x)\psi_j(x)^2 d\rho\\ &\quad+\sum_{\ell=n\tau+1}^{m}\int_X w(x)g(x)\psi_k(x)\frac{\sum_{j=1}^{n}\psi_j(x)^2}{n}d\rho\\ &= n\tau\int_X g(x)\psi_k(x)d\rho+(m-n\tau)\int_X g(x)\psi_k(x)d\rho=0,\end{aligned}$$

and this term again vanishes due to the orthogonality of g to ψ_k. For term II, splitting again the random samples we obtain

$$\begin{aligned}\sum_{k=1}^{n}\sum_{j=1}^{m}\mathbb{E}\left(\left(w(x^j)g(x^j)\psi_k(x^j)\right)^2\right) &= \sum_{j=1}^{m}\mathbb{E}\left(w(x^j)^2g(x^j)^2\sum_{k=1}^{n}\psi_k(x^j)^2\right)\\ &= n\sum_{j=1}^{m}\mathbb{E}\left(w(x)g(x)^2\right)\\ &= n\sum_{\ell=1}^{\tau}\sum_{j=1}^{n}\int_X w(x)g(x)^2\psi_j(x)^2 d\rho\\ &\quad+n\sum_{\ell=n\tau+1}^{m}\int_X w(x)g(x)^2\frac{\sum_{k=1}^{n}\psi_k(x)^2}{n}d\rho\\ &= n^2\tau\|g\|^2+n(m-n\tau)\|g\|^2=nm\|g\|^2.\end{aligned}$$

The proof of the last item is the same as the corresponding proof in Theorem 2, but using (2.16) with the random samples of Corollary 2. □

3. Adaptive approximation with a nested sequence of spaces.

We now apply the results for a given approximation space from the previous section to an arbitrary sequence of nested spaces $(V_k)_{k\geq 1}\subset L^2(X,\rho)$ with $V_k := \text{span}\{\psi_1,\dots,\psi_{n_k}\}$ and $n_k := \dim(V_k)$. Theorems 1 and 2 provide two different approaches to build the set of random samples for a given approximation space, and each one of them can be applied to the adaptive setting. Since the samples are adapted to the space, the underlying challenge is how to recycle as much as possible the samples associated to spaces from the previous iterations, in order to keep the overall number of generated samples from iteration one to k of the same order as $\dim(V_k)$, i.e., the same scaling as in the results for an individual approximation space.

First we briefly discuss the approach using Theorem 2. This theorem prescribes the precise number of random samples coming from each component of the mixture (2.4) associated to the space. When the spaces are nested, this trivially allows one to recycle all the samples from all the previous iterations, just by adding the missing samples to the previous ones, as shown in (1.2). The concrete procedure and the related Algorithm 1 are explained in section 4.1.

The approach using Theorem 1 is not as simple and effective as the previous one. Without recycling the samples from the previous iterations, the naïve sequential application of Theorem 1 to each space $V_1,\dots,V_t$ would require the generation of an overall number of samples equal to $\sum_{k=1}^{t}m_k$, with m_k samples drawn from each μ_{n_k}. However, despite μ_{n_k} changes at each iteration k, it is possible to recycle most, but not all, of the samples from the previous iterations by leveraging the additive structure of μ_{n_k} as in (1.3). This procedure is described in section 4.2 together with Algorithm 2.

The next results are obtained by applying Theorem 2 (respectively Theorem 1) individually for each space V_k and using a union bound, with the random samples produced by Algorithm 1 (respectively Algorithm 2). Here $I_k \in \mathbb{R}^{n_k \times n_k}$ denotes the identity matrix. For any $s > 1$, $\zeta(s)$ denotes the Riemann zeta function. The best approximation error (2.7) of u on the space V_k is denoted by $e_{n_k,2}(u)$, and u_C^k denotes the estimator (2.8) on V_k.

THEOREM 3. *Let $\alpha \in (0,1)$, $s > 1$, be real numbers and $t \geq 1$ be an integer. Given any nested sequence of spaces $V_1 \subset \cdots \subset V_t \subset L^2(X,\rho)$ with dimensions $n_1 < \cdots < n_t$, if*

$$m_k = \tau_k n_k, \quad \tau_k := \left\lceil \theta^{-1} \ln\left(\frac{\zeta(s)\, n_k^{s+1}}{\alpha}\right)\right\rceil, \qquad k = 1,\dots,t, \tag{3.1}$$

then

(i)
$$Pr\left(\bigcap_{k=1}^{t} \left\{ |||G_k - I_k||| \leq \frac{1}{2} \right\}\right) \geq 1 - \alpha,$$

where $G_k \in \mathbb{R}^{n_k \times n_k}$ is defined elementwise as

$$(G_k)_{pq} = m_k^{-1} \sum_{j=1}^{m_k} \psi_p\left(x^j\right) \psi_q\left(x^j\right),$$

and $x^1,\dots,x^{m_t}$ is a set of m_t independent random samples such that for any $k = 1,\dots,t$ and for any $j = 1,\dots,n_k$ the samples $x^{(j-1)\tau_k+1},\dots,x^{j\tau_k}$ are distributed according to χ_j. The set $x^1,\dots,x^{m_t}$ can be generated by Algorithm 1.

(ii) *If $u \in L^2(X,\rho)$, then for any $k = 1,\dots,t$ the estimator u_C^k satisfies,*

$$\mathbb{E}(\|u - u_C^k\|^2) \leq \left(1 + \frac{4\theta}{\ln(\zeta(s) n_k^{s+1}/\alpha)}\right) e_{n_k,2}(u)^2 + \alpha \|u\|^2.$$

Proof. Proof of (i). From Lemma 2, for any $k = 1,\dots,t$, Algorithm 1 with τ_k as in (3.1) generates a set $x^1,\dots,x^{m_k}$ of m_k random samples with the required properties. By construction, these random samples satisfy the assumptions of Theorem 2 and are used to compute the matrix G_k. For any $k = 1,\dots,t$, using Theorem 2 individually for each G_k with

$$\alpha_k = \frac{\alpha}{\zeta(s)\, n_k^s}$$

gives

$$\sum_{k=1}^{t} \Pr\left(\left\{ |||G_k - I_k||| > \frac{1}{2} \right\}\right) \leq \sum_{k=1}^{t} \alpha_k \leq \frac{\alpha}{\zeta(s)} \sum_{k=1}^{t} \frac{1}{k^s} \leq \frac{\alpha}{\zeta(s)} \sum_{k\geq 1} \frac{1}{k^s} = \alpha,$$

where the second inequality follows from strict monotonicity of $(n_k)_{k\geq 1}$ and $n_1 \geq 1$, which implies $n_k \geq k$.

Using De Morgan's law and a probability union bound for the matrices $G_1, \dots, G_t$ it holds that

$$\Pr\left(\bigcap_{k=1}^{t}\left\{|||G_k - I_k||| \le \frac{1}{2}\right\}\right) \ge 1 - \sum_{k=1}^{t} \Pr\left(\left\{|||G_k - I_k||| > \frac{1}{2}\right\}\right) \ge 1 - \alpha.$$

The proof of (ii) trivially follows from Theorem 2, since $\alpha_k \le \alpha$ for any $k = 1, \dots, t$. □

THEOREM 4. *Let $\alpha \in (0,1)$, $s > 1$, be real numbers and $t \ge 1$ be an integer. Given any nested sequence of spaces $V_1 \subset \cdots \subset V_t \subset L^2(X, \rho)$ with dimensions $n_1 < \cdots < n_t$, if*

$$m_k = \left\lceil \frac{n_k}{\theta} \ln\left(\frac{\zeta(s)\, n_k^{s+1}}{\alpha}\right)\right\rceil, \qquad k = 1, \dots, t, \tag{3.2}$$

then

(i)

$$Pr\left(\bigcap_{k=1}^{t}\left\{|||G_k - I_k||| \le \frac{1}{2}\right\}\right) \ge 1 - \alpha,$$

where $G_k \in \mathbb{R}^{n_k \times n_k}$ is defined elementwise as

$$(G_k)_{pq} = m_k^{-1} \sum_{j=1}^{m_k} \psi_p\left(x^j\right) \psi_q\left(x^j\right),$$

and $x^1, \dots, x^{m_t}$ is a set of m_t independent random samples such that for any $k = 1, \dots, t$ the $x^1, \dots, x^{m_k}$ are distributed according to μ_{n_k}. The set $x^1, \dots, x^{m_t}$ can be generated by Algorithm 2 using an overall number of random samples given by the random variable $\tilde{m}_t$ in (4.1)*.*

(ii) *If $u \in L^2(X, \rho)$, then for any $k = 1, \dots, t$ the estimator u_C^k satisfies,*

$$\mathbb{E}(\|u - u_C^k\|^2) \le \left(1 + \frac{4\theta}{\ln(\zeta(s) n_k^{s+1}/\alpha)}\right) e_{n_k,2}(u)^2 + \alpha \|u\|^2. \tag{3.3}$$

Proof. From Lemma 3, Algorithm 2 generates a set $x^1, \dots, x^{m_t}$ of m_t random samples with the required properties. The rest of the proof of items (i) and (ii) follows the proof of Theorem 3, but applying Corollary 1 individually to each G_k, rather than Theorem 2. □

Condition (3.1) ensures that m_k is an integer multiple of n_k for any $k \ge 1$. This condition requires a number of points m_k only slightly larger than condition (3.2) for the same values of n_k, s, and α (the proof is postponed to the upcoming Lemma 1). However, to compare the effective number of samples used in Theorems 3 and 4 one cannot just compare (3.1) and (3.2), because (3.2) neglects the random samples that have not been recycled from all the previous iterations. This issue is discussed in Remark 1.

For convenience in Remark 1, Lemma 1, and Remark 2 we denote with $\hat{m}_k$ the number of points required by condition (3.1) and with m_k the number of points required by (3.2), for the same values of n_k, s, and α.

Remark 1. For any $t \geq 1$, in Theorem 3 (Theorem 4) the set $x^1, \dots, x^{\hat{m}_t}$ ($x^1, \dots, x^{m_t}$) of random samples can be generated by Algorithm 1 (Algorithm 2). In Theorem 4, the generation of the m_t random samples requires Algorithm 2 to produce an overall number $\tilde{m}_t$ of random samples, with $\tilde{m}_t$ being the random variable defined in (4.1). Among the $\tilde{m}_t$ random samples (not necessarily distributed as μ_{n_t}) only m_t are retained, and the remaining $\tilde{m}_t - m_t$ are discarded. Condition (3.2) does not take into account the $\tilde{m}_t - m_t$ discarded samples. As a consequence, when comparing the effective number of samples required in Theorems 3 and 4, one should compare $\hat{m}_t$ with $\tilde{m}_t$, and not $\hat{m}_t$ with m_t.

It can be shown that $\tilde{m}_t - m_t$ remains small with large probability if m_k satisfies (3.2) for all $k = 1, \dots, t$. More precisely, using the upper bound $\tilde{m}_t \leq m_t + U_t$ with U_t defined in (4.2), Lemma 4, the last inequality in (3.4), and Remark 5, it can be shown that $\tilde{m}_t$ is upper bounded by a random variable with mean $(2+\theta)m_t$ and variance $(1+\theta)m_t$ that exhibits Gaussian concentration.

LEMMA 1. *For any $k \geq 1$ it holds that*

$$m_k \leq \hat{m}_k \leq m_k + n_k - 1 \leq m_k (1 + \epsilon_k) - 1, \tag{3.4}$$

where

$$\epsilon_k := \theta \left(\log \frac{2 n_k^{s+1} \zeta(s)}{\alpha} \right)^{-1} \ll 1.$$

Proof. For any $n_k \geq 1$ it holds that

$$\begin{aligned} m_k = \left\lceil n_k \theta^{-1} \log \frac{2 n_k^{s+1} \zeta(s)}{\alpha} \right\rceil &\leq n_k \left\lceil \theta^{-1} \log \frac{2 n_k^{s+1} \zeta(s)}{\alpha} \right\rceil \\ = \hat{m}_k &< n_k \left(\theta^{-1} \log \frac{2 n_k^{s+1} \zeta(s)}{\alpha} + 1 \right). \end{aligned}$$

The first inequality above proves the first inequality in (3.4). The rightmost strict inequality above and (3.2) prove the second (large) inequality in (3.4). The last inequality in (3.4) is obtained by using once again (3.2) together with properties of the ceiling operator. Notice that $\theta \approx 0.108$. ☐

Remark 2. The small number $\hat{m}_k - m_k < n_k$ of additional samples required by (3.1) contribute to further improve the stability of u_W. This slight surplus of samples is completely negligible from a fully adaptive point of view, where at each iteration k conditions (3.1) or (3.2) are not necessarily fulfilled, but, more simply, new random samples are just added to the previous ones until a certain stability criterion is met, for example, until $|||G_k - I_k||| \leq \xi_k$ or $\text{cond}(G_k) \leq \xi_k$ for some threshold ξ_k eventually depending on k.

Remark 3. For any integer $t \geq 1$ and reals $\alpha \in (0,1)$, $s > 1$, the result in Theorem 3 can be sharpened by using m_t random samples $x^1, \dots, x^{m_t}$ such that, with the same τ_k as in (3.1),

- for any $k = 1, \dots, t-1$ and for any $j = 1, \dots, n_k$ the $x^{(j-1)\tau_k + 1}, \dots, x^{j\tau_k}$ are distributed according to χ_j,
- at iteration t, for any $j = 1, \dots, n_t$ the samples $x^{(j-1)\tau_{t-1}+1}, \dots, x^{j\tau_{t-1}}$ are distributed according to χ_j, and the samples $x^{n_t \tau_{t-1}+1}, \dots, x^{m_t}$ are distributed according to μ_{n_t}.

For any $k = 1, \dots, t-1$ the set $x^1, \dots, x^{m_k}$ can be incrementally constructed using (1.2). For the construction of $x^1, \dots, x^{m_t}$ at the last iteration t: we recycle all the

$m_{t-1} = n_{t-1}\tau_{t-1}$ samples from iteration $t-1$, then we add $(n_t - n_{t-1})\tau_{t-1}$ new samples, i.e., τ_{t-1} samples from χ_j for all $j = n_{t-1}+1, \dots, n_t$, and then add the remaining $m_t - n_t\tau_{t-1}$ samples from μ_{n_t}.

Using the random samples $x^1, \dots, x^{m_t}$, in the proof of Theorem 3 we can apply Theorem 2 at the first $t-1$ iterations, and then at iteration t apply Corollary 2 with $n = n_t$, $m = m_t$, $\tau = \tau_{t-1}$. This proves the same conclusions of Theorem 3 under the same condition (3.1) on m_k for $k = 1, \dots, t-1$, but with the slightly better condition (3.2) on m_t, because m_t need not be an integer multiple of n_t.

Remark 4. Remark 3 can be used to develop adaptive algorithms that recycle all the samples from all the previous iterations, and that use a number of samples m_t given by (3.2) at the last iteration t. Such adaptive algorithms need to detect in advance which is last iteration, in contrast to algorithms developed using Theorem 3 where this information is not needed.

4. Sampling algorithms.

In the following we present two sequential algorithms that generate the random samples required by Theorem 3 or Theorem 4 at any iteration, say t, while recycling the samples from the previous iterations $k = 1, \dots, t-1$. The algorithms are described in the appendix, using the convention that a loop **for** $i = \mathit{start}$ **to** end on the variable say i is not executed if $\mathit{end} < \mathit{start}$.

4.1. Deterministic sequential sampling.

This section presents Algorithm 1, which can be used to produce the random samples required by Theorem 3 using the decomposition (1.2). By construction, Algorithm 1 recycles all the samples from all the previous iterations. At any iteration $t \geq 1$ the algorithm stores the $m_t = \tau_t n_t$ random samples in an $n_t \times \tau_t$ matrix. All the elements of this matrix are modified only once as the algorithm runs from iteration one to t. The algorithm works with any nondecreasing positive integer sequence $(\tau_k)_{k\geq 1}$.

LEMMA 2. *Let $(V_k)_{k\geq 1}$ be any sequence of nested spaces with dimension $n_k = \dim(V_k)$, and $(\tau_k)_{k\geq 1}$ be a positive nondecreasing integer sequence. For any $t \geq 2$, Algorithm* 1 *generates a set of $m_t = \tau_t n_t$ random samples $x^1, \dots, x^{m_t}$ with the property that for any $k = 1, \dots, t$ and for any $j = 1, \dots, n_k$ the samples $x^{(j-1)\tau_k+1}, \dots, x^{j\tau_k}$ are distributed according to χ_j.*

Proof. At any iteration $k = 1, \dots, t$ Algorithm 1 produces the matrix $\{x^{j\ell}, j = 1, \dots, n_k, \ell = 1, \dots, \tau_k\}$ that contains the $\tau_k n_k = m_k$ random samples, by modifying only the elements $\{x^{j\ell}, j = 1, \dots, n_{k-1}, \ell = 1 + \tau_{k-1}, \dots, \tau_k\}$ and $\{x^{j\ell}, j = 1 + n_{k-1}, \dots, n_k, \ell = 1, \dots, \tau_k\}$. By construction, for any $j = 1, \dots, n_k$, the jth row of this matrix contains τ_k samples distributed as χ_j. This matrix is recast into a column vector by means of a transposition composed with a vectorization, which piles up its rows into the vector $(x^1, \dots, x^{m_k})^\top$. □

When ρ is a product measure on X, random samples from all the χ_j appearing in Algorithm 1 can be efficiently drawn by using the algorithms proposed in [7], i.e., *inverse transform sampling* or *rejection sampling*. The computational cost required by these algorithms scales linearly with respect to d and to the desired number of samples.

4.2. Random sequential sampling.

This section presents Algorithm 2, which can be used to produce the random samples required by Theorem 4, and uses the decomposition (1.3). For any $k \geq 2$, a standard algorithm for generating m_k random samples from μ_{n_k} uses a binomial random variable $B_k \sim \mathrm{Bin}(m_k, (n_k - n_{k-1})/n_k)$ to determine the proportion of samples coming from σ_{n_k}. The first parameter of

B_k is the number of trials, and the second parameter is the probability of success for each trial, that is given by the coefficient multiplying $d\sigma_{n_k}$ in (1.3). For any $k \neq k'$, B_k and $B_{k'}$ are mutually independent. The amount of samples associated to $\mu_{n_{k-1}}$ is equal to $m_k - B_k$. These are the samples that the algorithm can recycle from the previous iterations, whenever necessary. For any $t \geq 1$, the algorithm that generates random samples from $\mu_{n_1}, \dots, \mu_{n_t}$ in a sequential manner is described in Algorithm 2. Efficient algorithms have been proposed in [7] for drawing samples from all the probability measures μ_{n_k} and σ_{n_k} appearing in Algorithm 2. The next lemma quantifies more precisely how many unrecycled samples cumulate after say t iterations.

LEMMA 3. *For any $t \geq 1$, Algorithm 2 generates a set of m_t random samples $x^1, \dots, x^{m_t}$ with the property that $x^1, \dots, x^{m_k}$ are distributed according to μ_{n_k}, for any $k = 1, \dots, t$. The overall number of samples generated by Algorithm 2 at iteration t is*

$$\tilde{m}_t := m_1 + \sum_{k=2}^{t} \left(B_k + \max\{m_k - B_k - m_{k-1}, 0\} \right). \tag{4.1}$$

Proof. The properties of the random samples are ensured by construction. We now proove (4.1). When $t = 1$ the sum is empty and the formula holds true. Suppose then $t \geq 2$. The proof uses induction on k. At iteration $k = 2$, m_{k-1} samples from $\mu_{n_{k-1}}$ are available, which verifies the induction hypothesis. The proof of the induction step is as follows. For any $k \geq 2$, supposing that m_{k-1} samples from $\mu_{n_{k-1}}$ are available at iteration $k-1$, the number of recycled samples from iteration $k-1$ is $\min(m_k - B_k, m_{k-1})$. Then the algorithm adds $\max(m_k - B_k - m_{k-1}, 0)$ new samples from $\mu_{n_{k-1}}$. Afterward $m_k - \max(m_k - B_k - m_{k-1}, 0) - \min(m_k - B_k, m_{k-1})$ new samples are added from σ_{n_k}. At the end of iteration k, the algorithm produces a set containing m_k random samples from μ_{n_k} and throws away $m_{k-1} - \min(m_k - B_k, m_{k-1})$ samples that were drawn at iteration $k-1$ from $\mu_{n_{k-1}}$. Summation of each contribution of new samples at any iteration k from 2 to t gives (4.1). ∎

The number of unrecycled samples after t iterations is $\tilde{m}_t - m_t$. As a sum of nonnegative random variables, this number can only increase as the algorithm runs, which represents the major disadvantage of Algorithm 2, and of any other purely random sequential algorithm. Since $m_k \geq m_{k-1}$ and $B_k \geq 0$ for all $k \geq 2$, from (4.1) we have the upper bound $\tilde{m}_t \leq m_t + U_t$, where U_t is the random variable defined as

$$U_t := \sum_{k=2}^{t} B_k \tag{4.2}$$

that gives an upper bound for the number of unrecycled samples. Its mean and variance are given by

$$\mathbb{E}(U_t) = \sum_{k=2}^{t} \mathbb{E}(B_k) = \sum_{k=2}^{t} m_k \frac{n_k - n_{k-1}}{n_k}, \quad \mathrm{Var}(U_t) = \sum_{k=2}^{t} \mathrm{Var}(B_k) = \sum_{k=2}^{t} m_k \frac{n_k - n_{k-1}}{n_k} \frac{n_{k-1}}{n_k}. \tag{4.3}$$

The above expressions for the mean and variance of U_t hold true for any condition between m_k and n_k, not necessarily of the form (3.2). When (3.2) is fulfilled we have the following upper bounds.

LEMMA 4. *For any strictly increasing sequence* $(n_k)_{k\geqslant 1}$, *for any* $t \geqslant 2$, $s > 1$, *and* $\alpha \in (0,1)$, *if* m_k *and* n_k *satisfy condition* (3.2) *for all* $k = 1, \dots, t$, *then*

$$\mathit{Var}(U_t) < \mathbb{E}(U_t) \leqslant m_t + n_t - m_1.$$

Proof.

$$\begin{aligned}
\mathbb{E}(U_t) &= \sum_{k=2}^{t} m_k \frac{n_k - n_{k-1}}{n_k} \\
&\leq \sum_{k=2}^{t} (n_k - n_{k-1}) \left\lceil \theta^{-1} \ln\left(\frac{\zeta(s)\, n_k^{s+1}}{\alpha} \right) \right\rceil \\
&\leq \sum_{k=2}^{t} n_k \left\lceil \theta^{-1} \ln\left(\frac{\zeta(s)\, n_k^{s+1}}{\alpha} \right) \right\rceil - n_{k-1} \left\lceil \theta^{-1} \ln\left(\frac{\zeta(s)\, n_{k-1}^{s+1}}{\alpha} \right) \right\rceil \\
&= n_t \left\lceil \theta^{-1} \ln\left(\frac{\zeta(s)\, n_t^{s+1}}{\alpha} \right) \right\rceil - n_1 \left\lceil \theta^{-1} \ln\left(\frac{\zeta(s)\, n_1^{s+1}}{\alpha} \right) \right\rceil \\
&\leq m_t + n_t - m_1.
\end{aligned}$$

The inequality for $\mathrm{Var}(U_t)$ follows from (4.3) and strict monotonicity of the sequence $(n_k)_k$. □

Remark 5. Here we show that the random variable U_t concentrates like a Gaussian random variable with mean and variance given by (4.3). The central limit theorem for a binomial random variable $B \sim \mathrm{Bin}(m,p)$ with number of trials m and success probability p states that

$$\lim_{m\to\infty} \Pr\left(\frac{B - mp}{\sqrt{mp(1-p)}} \leq b \right) = \Phi\left(b\right), \quad b \in \mathbb{R},$$

where Φ is the cumulative distribution function of the standard Gaussian distribution. This justifies the well-known Gaussian approximation of B when m is sufficiently large. This approximation is very accurate already when $mp \geq 5$ and $m(1-p) \geq 5$. In our settings, when m_k and n_k satisfy (3.2) for some $\alpha \in (0,1)$ and $s > 1$, the parameters of the binomial random variables B_k overwhelmingly verify the above conditions for any $k = 2, \dots, t$, since

$$\frac{m_k(n_k - n_{k-1})}{n_k} \geq \theta^{-1} \ln\left(\frac{\zeta(s)\, n_k^{s+1}}{\alpha} \right)(n_k - n_{k-1}) \geq \theta^{-1} \ln\left(\frac{\zeta(s)\, n_k^{s+1}}{\alpha} \right) \gg 5, \tag{4.4}$$

$$\frac{m_k n_{k-1}}{n_k} = \theta^{-1} \ln\left(\frac{\zeta(s)\, n_k^{s+1}}{\alpha} \right) n_{k-1} \gg 5, \tag{4.5}$$

and $\theta^{-1} \approx 9.242$. Using the Gaussian approximation of the binomial distribution, each B_k behaves like a Gaussian random variable with the same mean and variance. A finite linear combination of independent Gaussian random variables is a Gaussian random variable. Hence the random variable U_t behaves like a Gaussian random variable with mean and variance as in (4.3).

4.3. Comparison of the sampling algorithms. The main properties of Algorithms 1 and 2 are resumed below. At any iteration, say t, the following hold.

- Algorithm 1 generates $m_t = \tau_t n_t$ independent random samples with τ_t being any positive integer and such that τ_t of these random samples are drawn from χ_j for any $j = 1, \dots, n_t$. This algorithm recycles all the samples generated at all the previous iterations $k = 1, \dots, t-1$.
- Algorithm 2 generates m_t independent random samples from μ_{n_t}. This algorithm recycles most of the samples generated at all the previous iterations $k = 1, \dots, t-1$. If (3.2) holds true at any iteration, then the number of unrecycled samples at iteration t is upper bounded by the random variable (4.2) with mean $(1+\theta)m_t$ and variance $(1+\theta)m_t$, which exhibits Gaussian concentration.
- During the execution, Algorithm 1 modifies each element of the output set $x^1, \dots, x^{m_t}$ only once, in contrast to Algorithm 2, which can modify the same element several times, when discarding previously generated random samples.
- Algorithms 1 and 2 use any sequence of nested spaces $(V_k)_k$.
- The WLS estimators constructed with the random samples generated by both Algorithms 1 and 2 share the same theoretical guarantees; see Theorems 3 and 4.
- In practice Algorithm 1 outperforms Algorithm 2 in all our numerical tests, recycling all the samples from all the previous iterations, and producing on average more stable Gramian matrices.

Remark 6. When using random samples from γ^m rather than from μ^m, the benefits of variance reduction increase with more localized basis than orthogonal polynomials, like wavelets. The structure of the random samples from γ^m ensures that for any element of the basis $\psi_j \in V$ at least one sample is contained in $\text{supp}(\psi_j)$. If this is not the case, then the Gramian matrix is singular, because the discrete inner product of two functions is equal to zero when none of the samples is contained in the intersection of their supports.

5. Numerical methods for adaptive (polynomial) approximation.

The results presented in Theorems 3 and 4 hold for any nested sequence $(V_k)_k$ of general approximation spaces, in any dimension d. Two families of spaces that are suitable for approximation in arbitrary dimension d are polynomial spaces and wavelet spaces. In this paper we confine ourselves to polynomial spaces. Even with this restriction, adaptive numerical methods in such a general context are still quite a large subject. Our focus in the present paper is on a more specific type of adaptive methods, in the spirit of orthogonal matching pursuit, and along the line of the greedy algorithms described in [10].

The spaces V_k can be adaptively chosen from one iteration to the other, as long as the sequence remains nested. Without additional information on the function that we would like to approximate, the infinite number of elements in the basis prevents the development of a concrete strategy for performing the adaptive selection. Such additional information is available in the form of decay of the coefficients, for example, for some PDEs with parametric or stochastic coefficients, whose solution is provably well-approximated by so-called downward closed polynomial spaces. See [6] and references therein for an introduction to the topic. The definition of downward closed polynomial spaces is postponed to (5.3). In the rest of this section we assume that

$$\begin{gathered}\text{the function } u \text{ can be well approximated by a nested sequence}\\ \text{of downward closed polynomial approximation spaces.}\end{gathered} \tag{5.1}$$

As a relevant example that motivates our interest in the above setting, for PDEs with lognormal diffusion coefficients it was shown by the author in [8, Lemma 2.4] that suitable polynomial spaces yielding provable convergence rates are actually downward closed.

After (5.1) we restrict our analysis to nested sequences $(V_k)_k$ of polynomial spaces satisfying the additional constraint of being downward closed. At iteration k, given V_{k-1}, an ideal (local) optimal criterion for performing the adaptive selection is to choose $V_k \supset V_{k-1}$ as the space that delivers the smallest error among all possible downward closed spaces with prescribed dimension, for example $n_k = 1 + n_{k-1}$. Since d is finite, the number of all possible choices for V_k is also finite. In reality the exact error $\|u - \Pi_{n_k} u\|$ is not available, and the adaptive selection has to rely on the error $\|u - u_C^k\|$ that is a random variable. Here the error estimates from Theorems 3 and 4 come in handy because they ensure that $\|u - u_C^k\|^2$ is less than twice $\|u - \Pi_{n_k} u\|^2$ in expectation. Even if the exact error was available, the adaptive selection using the local optimal criterion does not ensure optimality of the selected spaces at the following iterations, and for this reason it is referred to as a *greedy* adaptive selection.

Before moving to the description of the adaptive algorithm, we briefly introduce some definitions that are useful to describe the polynomial setting. Hereafter we assume that $X = \times_{i=1}^d I_i$ is the Cartesian product of intervals $I_i \subset \mathbb{R}$ and that $d\rho = \otimes_{i=1}^d d\rho_i$, where each ρ_i is a probability measure defined on I_i. This setting ensures the existence of a product basis orthonormal in $L^2(X, \rho)$ that we now introduce. To simplify the presentation and notation, we further suppose that $I := I_j$ and $\tilde{\rho} := \rho_j$ for any j, and denote with $(T_j)_{j \geq 1}$ the univariate family of orthogonal polynomials, orthonormal in $L^2(I, \tilde{\rho})$. Let $\Lambda \subset \mathcal{F} := \mathbb{N}_0^d$ be a multi-index set enumerated according to an ordering relation, for example, the lexicographical ordering. Using Λ we define

$$\psi_\nu(x) := \prod_{i=1}^d T_{\nu_i}(x_i), \quad \nu = (\nu_1, \dots, \nu_d) \in \Lambda, \quad x = (x_1, \dots, x_d) \in X, \tag{5.2}$$

and relate the orthonormal basis $(\psi_i)_{i \geq 1}$ from the previous sections to the above orthonormal basis as $\psi_i = \psi_{\nu^i}$ for any $i = 1, \dots, \#(\Lambda)$, where ν^i is the ith element of Λ according to the lexicographical ordering, and $\#(\Lambda)$ denotes the cardinality of Λ. The space associated to Λ is defined as $V_\Lambda := \operatorname{span}\{\psi_\nu : \nu \in \Lambda\}$.

A set $\Lambda \subset \mathcal{F}$ is downward closed if

$$\nu \in \Lambda \text{ and } \nu' \leq \nu \implies \nu' \in \Lambda, \tag{5.3}$$

where the ordering $\nu' \leq \nu$ is intended in the lexicographical sense. We say that the space V_Λ is downward closed if the supporting index set Λ is downward closed. For any $\Lambda \subset \mathcal{F}$ downward closed we define its margin $\mathcal{M}(\Lambda)$ as

$$\mathcal{M}(\Lambda) := \{\nu \in \mathcal{F} : \nu \notin \Lambda \wedge \exists j \in \{1, \dots, d\} : \nu - e_j \in \Lambda\},$$

where $e_j \in \mathcal{F}$ is the multi-index with all components equal to zero, except the jth component that is equal to one. The reduced margin $\mathcal{R}(\Lambda)$ of Λ is defined as

$$\mathcal{R}(\Lambda) := \{\nu \in \mathcal{F} : \nu \notin \Lambda \wedge \forall j \in \{1, \dots, d\}, \nu_j \neq 0 \implies \nu - e_j \in \Lambda\} \subseteq \mathcal{M}(\Lambda).$$

If Λ is downward closed, then $\Lambda \cup \{\nu\}$ is downward closed for any $\nu \in \mathcal{R}(\Lambda)$.

Finally we choose the space V_k from the previous sections as $V_k = V_{\Lambda_k}$ for any k by means of a nested sequence $(\Lambda_k)_k \subset \mathcal{F}$ of downward closed multi-index sets. For any $k \geq 1$, $\#(\Lambda_k) = \dim(V_k) = n_k$ equals the dimension of V_k.

5.1. An adaptive orthogonal matching pursuit algorithm. In this section we describe an adaptive algorithm using optimal weighted least squares, starting from the algorithm proposed in [18] for standard least squares and inspired by the orthogonal matching pursuit. The algorithm builds a sequence of nested spaces $V_{\Lambda_1} \subset \cdots \subset V_{\Lambda_t}$ performing at each iteration an adaptive greedy selection of the indices identifying the elements of the basis. The adaptive construction of the index sets uses ideas that were originally proposed in [14] for developing adaptive sparse grids quadratures. The greedy selection of the indices uses a marking strategy known as bulk chasing.

The adaptive algorithm works with downward closed index sets. Given any Λ downward closed, a nonnegative function $e : \mathcal{R}(\Lambda) \to \mathbb{R}$, and a parameter $\beta \in (0, 1]$, we define the procedure $\text{BULK} := \text{BULK}(\mathcal{R}(\Lambda), e, \beta)$ that computes a set $F \subseteq \mathcal{R}(\Lambda)$ of minimal positive cardinality such that

$$\sum_{\nu \in F} e(\nu) \geq \beta \sum_{\nu \in \mathcal{R}(\Lambda)} e(\nu). \tag{5.4}$$

Denote with a_ν the coefficient associated to ψ_ν in the expansion $u = \sum_{\nu \in \mathcal{F}} a_\nu \psi_\nu$. For any $\nu \in \mathcal{R}(\Lambda)$, the function $e(\nu)$ is chosen as an estimator for $|a_\nu|^2$. The adaptive algorithm is described in Algorithm 3. At any iteration k of the algorithm, m_k random samples are generated by using Algorithm 1, with m_k satisfying (3.1) as a function of $n_k = \#(\Lambda_k)$, for a given choice of the parameters α and s. The m_k random samples are used to compute the weighted least-squares estimator u_C^k on V_{Λ_k}. For convenience in Algorithm 3 the operations performed by Algorithm 1 have been merged with those for the adaptive selection of the space. In Algorithm 3 the χ_ν correspond to χ_j with $\psi_j = \psi_\nu$. An estimator for $|a_\nu|^2$ proposed in [18] that uses only the information available at iteration $k - 1$ is

$$e_{k-1}(\nu) := \left| \langle u - u_C^{k-1}, \psi_\nu \rangle_{m_{k-1}} \right|^2, \quad \nu \in \mathcal{R}(\Lambda_{k-1}), \tag{5.5}$$

where the discrete inner product uses the evaluations of the function u at the same m_{k-1} samples that have been used to compute u_C^{k-1} at iteration $k-1$. The estimator (5.5) uses the residual $r_{k-1} := u - u_C^{k-1}$ and is cheap to compute: it requires only the product of a vector with a matrix.

A safeguard mechanism prevents Algorithm 3 from getting stuck into indices associated to null coefficients in the expansion of u_C^k. Given a positive integer k_{sg}, once every k_{sg} iterations the algorithm adds to Λ_k the most ancient multi-index from $\mathcal{R}(\Lambda_{k-1}) \setminus F$. In the numerical tests reported in the next section, such a mechanism was never activated, and the algorithm was always able to identify the best n_k-term index sets of the given function at any iteration k.

Algorithm 3 can be modified by relaxing (3.1) to a less demanding condition between m_k and n_k at each iteration k. For example, the random samples can be added until a stability condition of the form $|||G_k - I_k||| \leq \xi$ is met, for some given threshold $\xi > 1/2$. This provides a fully adaptive algorithm as described in Algorithm 4 that however, in contrast to Algorithm 3, does not come with the theoretical guarantees of Theorem 3.

5.2. Testing the sampling algorithms. This section presents some numerical tests of the sampling algorithms that generate the random samples, comparing Algorithms 1 and 2. At the very end, our implementation of both algorithms uses inverse transform sampling as described in [7, section 5.2] for drawing samples from all the χ_j.

A natural vehicle to quantify the quality of the generated samples is the deviation of the matrix G_k from the identity, i.e., $|||G_k - I_k|||$. Since $|||G_k - I_k||| \leq \frac{1}{2} \implies$

$\mathrm{cond}(G_k) := |||G_k^{-1}||| \, |||G_k||| \leq 3$, our tests show the condition number, which is a more meaningful quantity when solving a linear system.

From the point of view of the stability and convergence properties of the WLS estimators, the random samples generated by both algorithms come with the same theoretical guarantees. But, in contrast to Algorithm 2, Algorithm 1 recycles all the samples from the previous iterations. This is the main reason to prefer Algorithm 1 over Algorithm 2. Another reason to choose Algorithm 1 is that it produces more stable Gramian matrices on average, since the sample variance of the generated samples is lower.

Our first tests illustrate the benefits of variance reduction, using spaces V_k of univariate Hermite polynomials $(H_j)_{j\geq 0}$ with degree from 0 to $k-1$. More precisely, the sequence $(H_j)_{j\geq 0}$ contains univariate Hermite polynomials orthonormalized as $\int_{\mathbb{R}} H_i(t)H_j(t)\,dg = \delta_{ij}$, where $dg := (2\pi)^{-1/2}e^{-t^2/2}\,dt$. Denote with $E_i \approx \mathbb{E}(\mathrm{cond}(G_k))$ and $S_i^2 \approx \mathrm{Var}(\mathrm{cond}(G_k))$ the sample mean and sample variance estimators of the random variable $\mathrm{cond}(G_k)$ with G_k constructed using the random samples generated by Algorithm $i \in \{1,2\}$. Figure 1, left, shows the comparison of E_i and $E_i + S_i$ between the two algorithms, with $m_k = \lceil \theta^{-1} \rceil n_k$. Both estimators confirm that Algorithm 1 produces random samples whose Gramian matrix is better conditioned than Algorithm 2. The same trend persists when choosing other scalings like $m_k = (3+n_k)n_k$; see Figure 1, center and right. The difference between the two algorithms is expected to amplify when using more localized bases, with Algorithm 2 producing more ill-conditioned Gramian matrices as the ratio m_k/n_k decreases.

From now on the focus is on Algorithm 1. For all the tests in the remaining part of this section we choose m_k as in (3.1) with $\alpha = 0.1$ and $s = 2$. The value of α is chosen fairly large on purpose to check, in practice, how sharp the stability constraint (3.1) is. In the first test, we choose $\rho = \otimes^d dg$ as the d-dimensional probabilistic Gaussian measure on $X = \mathbb{R}^d$ and V_k as the spaces of tensorized Hermite polynomials, obtained from (5.2) by taking $T_j = H_j$, $j \geq 0$. The Gaussian case poses several challenges: as shown in [7], standard least-squares estimators with Hermite polynomials typically fail due to the ill-conditioning of the Gramian matrix. Since the ill-conditioning arises with high-degree polynomials, we choose fairly low-dimensional tests to begin with, such that very-high-degree polynomials can be tested, e.g., degrees beyond 100. With $d = 1$ the results are shown in Figure 2, left, and with $d = 4$ in Figure 2, right. The condition number of G_k stays well below the threshold equal to 3 during all the simulations, which contain, respectively, 10^4 and 10^3 realizations of the sequence $(\mathrm{cond}(G_k))_{k\geq 1}$ with random samples generated by Algorithm 1. At

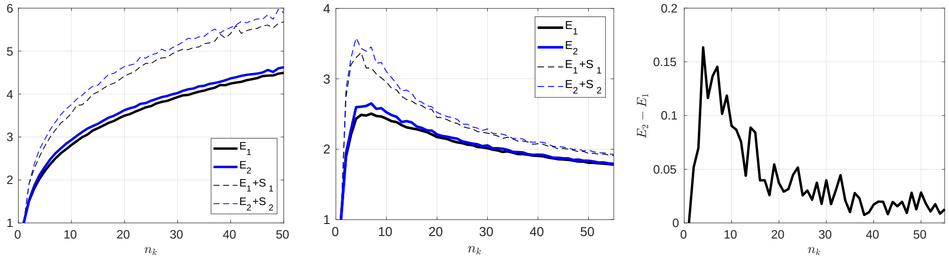

FIG. 1. *Left: Estimators E_i and $E_i + S_i$ of the sequence of random variables $(cond(G_k))_{k\geq 1}$ at iteration $k = 1,\ldots,50$ with $n_k = k$ and $m_k = \lceil\theta^{-1}\rceil n_k$. Hermite polynomials. $d = 1$. The estimators use 10^4 realizations of the sequence $(cond(G_k))_{k\geq 1}$. Center: Estimators E_i and $E_i + S_i$ of the sequence of random variables $(cond(G_k))_{k\geq 1}$ at iteration $k = 1,\ldots,55$ with $n_k = k$ and $m_k = (3+n_k)n_k$. Hermite polynomials. $d = 1$. The estimators use 10^3 realizations of the sequence $(cond(G_k))_{k\geq 1}$. Right: Same simulation as center but showing $E_2 - E_1$.*

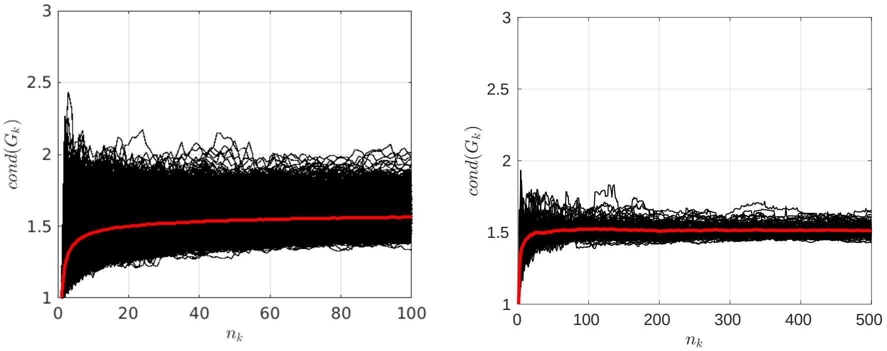

FIG. 2. *Left: Condition number $cond(G_k)$ at iteration k with $n_k = k$ and m_k as in* (2.15)*, $d = 1$, Gaussian measure, Hermite polynomials, $s = 2$, $\alpha = 0.1$. Black lines are 10^4 realizations of the sequence $(cond(G_k))_{k\geq 1}$ with random samples from Algorithm* 1*. The red line is their sample mean. Right: Condition number $cond(G_k)$ at iteration k with $n_k = k$ and m_k as in* (2.15)*, $d = 4$, Gaussian measure, Hermite polynomials, $s = 2$, $\alpha = 0.1$. Black lines are 10^3 realizations of the sequence $(cond(G_k))_{k\geq 1}$ with random samples from Algorithm* 1*. The red line is their sample mean.*

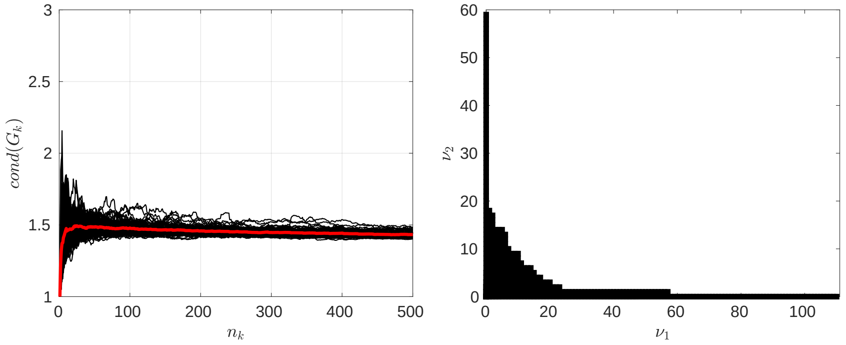

FIG. 3. *Left: Condition number $cond(G_k)$ at iteration k with $n_k = k$ and m_k as in* (2.15)*, $d = 4$, uniform measure, Legendre polynomials, $s = 2$, $\alpha = 0.1$. Black lines are 10^3 realizations of the sequence $(cond(G_k))_{k\geq 1}$ with random samples from Algorithm* 1*. The red line is their sample mean. Right: Section of the first and second coordinates of an index set obtained at iteration $k = 500$ during the simulation in Figure* 2*, right.*

any iteration k, the index set $\Lambda_k \supset \Lambda_{k-1}$ that defines the space $V_k = V_{\Lambda_k}$ is generated by adding to Λ_{k-1} a random number of indices randomly chosen from $\mathcal{R}(\Lambda_{k-1})$. This procedure generates nested sequences of downward closed index sets; see Figure 3, right, for an example of such a set. With other families of orthogonal polynomials the results are very similar. For example, with $d = 4$, the results in Figure 3, left, with the d-dimensional uniform probabilistic measure on $X = [-1, 1]^d$ and Legendre polynomials are analogous to those obtained in Figure 2, right, with the Gaussian measure and Hermite polynomials. Figure 3, right, shows an example of (the section of the first and second coordinates of) an index set Λ_k obtained in the simulation of Figure 2, right, at iteration $k = 500$. This set contains products of univariate Hermite polynomials with degree over 110 in the first coordinate and up to 59 in the second coordinate, and degree up to 25 and 9 in the remaining third and fourth coordinates not displayed in the figure.

5.3. Testing the adaptive algorithm. For the numerical tests of Algorithm 3 we choose ρ as the uniform measure over $X = [-1, 1]^d$ and V_k as the spaces of tensorized Legendre polynomials obtained by first defining the sequence $(L_j)_{j\geq 0}$ of

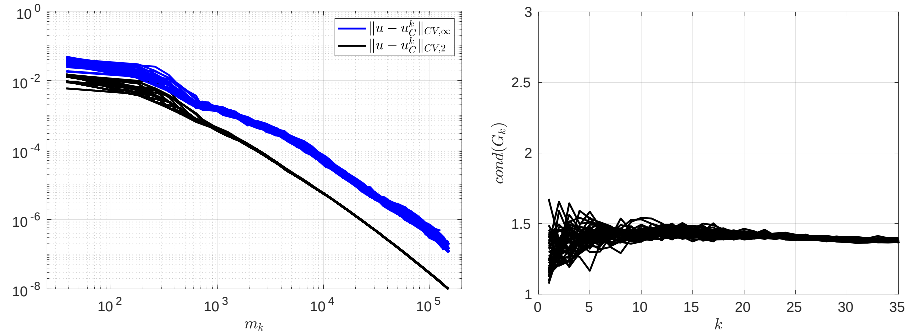

FIG. 4. *Left:* 10^2 *realizations of the errors* $\|u-u_C^k\|_{CV,2}$ *and* $\|u-u_C^k\|_{CV,\infty}$ *versus* m_k *obtained with Algorithm* 3 *and the random samples generated by Algorithm* 1. *Right:* 10^2 *realizations of* $cond(G_k)$ *versus* k*, for the same simulation on the left.*

univariate Legendre polynomials orthonormalized as $\int_{-1}^{+1} L_i(t)L_j(t)\,\frac{dt}{2} = \delta_{ij}$ and then taking $T_j = L_j$ in (5.2). As an illustrative example, consider the following function that satisfies assumption (5.1):

$$u(x) = \left(1 + \frac{1}{2d}\sum_{i=1}^{d} q_i x_i\right)^{-1}, \qquad x \in X, \tag{5.6}$$

with $d = 16$ and $q_i = 10^{-\frac{3(i-1)}{d-1}}$. A set X_{CV} of 10^6 cross-validation points uniformly distributed over X is chosen once and for all, and the approximation error $\|u - u_C^k\|$ is estimated as

$$\begin{aligned} \|u-u_C^k\| \approx \|u-u_C^k\|_{CV,2} &:= \sqrt{\frac{1}{\#(X_{CV})}\sum_{\tilde{x}\in X_{CV}} |u(\tilde{x}) - u_C^k(\tilde{x})|^2} \\ &\leq \|u-u_C^k\|_{CV,\infty} := \max_{\tilde{x}\in X_{CV}} |u(\tilde{x}) - u_C^k(\tilde{x})|. \end{aligned} \tag{5.7}$$

The error estimators are denoted with $\|u-u_C^k\|_{CV,2}, \|u-u_C^k\|_{CV,\infty}$, although these are not norms over the functional space. The parameter of the marking strategy is set to $\beta = 0.5$, and $\Lambda_1 = \{(0,\dots,0)^\top\}$. Figure 4, left, shows the results for the errors (5.7) obtained when approximating the function (5.6) with Algorithm 3 and using the random samples generated by Algorithm 1. At each iteration k the number of samples m_k as a function of n_k satisfies (3.1) with $\alpha = 0.1$ and $s = 2$. Figure 4, right, shows the condition number of G_k at iteration k, which stays below two at all the iterations. Figure 5, left, shows that at each iteration k the adaptive algorithm catches the coefficients in the best n_k-term set. The coefficients in Figure 5, left, have not been sorted, and they appear in the same order in which their corresponding elements of the basis were included in the approximation space by the adaptive selection procedure. After 35 iterations the algorithm has adaptively constructed a sequence $\Lambda_1, \dots, \Lambda_{35}$ of index sets. The set Λ_{35} contains about 10^3 indices, and its associated space $V_{\Lambda_{35}}$ provides an approximation error of the order 10^{-7} on average. Figure 5 shows some sections of Λ_{35}. All the d coordinates in Λ_{35} are active, i.e., for all $i \in \{1,\dots,d\}, \exists\nu \in \Lambda : \nu_j > 0$.

The condition number in Figure 4, right, actually decreases w.r.t. k, showing that condition (3.1) could be relaxed while still preserving the stability of the discrete projection, and yielding faster convergence rates w.r.t. m_k than those in Figure 4, left.

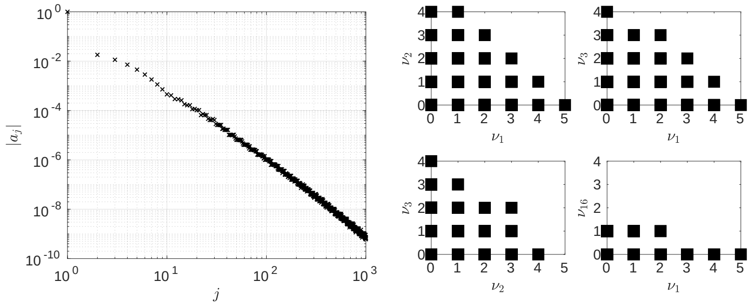

Fig. 5. *Left: First* 10^3 *coefficients of the estimator* $u_C^k = \sum_j a_j \psi_j$ *obtained at iteration* $k = 35$ *with the index set* Λ_{35}*, for one realization among those shown in Figure* 4*. Right: Some sections of the index set corresponding to the coefficients displayed on the left.*

6. Conclusions. We have advanced one step further the analysis of optimal WLS estimators for a given general d-dimensional approximation space. The main novelty concerns the structure of the random samples, which follow a distribution with product form. The results have immediate applications to the adaptive setting with a nested sequence of approximation spaces and point out new promising directions for the development of adaptive numerical methods for high-dimensional approximation using polynomial or wavelet spaces. Our analysis indicates that efficient adaptive methods can also be developed for general sequences of nonnecessarily nested spaces. This topic will be investigated in the future.

Appendix A. Algorithms.

Algorithm 1. Deterministic sequential sampling.

```
INPUT: $t$, $d\rho$, $(\tau_k)_{k=1}^t$, $(n_k)_{k=1}^t$, $(\psi_j)_{j=1}^{n_t}$
OUTPUT: $x^1, \dots, x^{m_t}$ s.t. $x^{(j-1)\tau_k+1}, \dots, x^{j\tau_k} \overset{\text{i.i.d.}}{\sim} \chi_j, j = 1, \dots, n_k, k = 1, \dots, t$.
  for $j = 1$ to $n_1$ do
    for $\ell = 1$ to $\tau_1$ do
      Sample $x^{j\ell}$ from $\chi_j$
    end for
  end for
  $(x^1, \dots, x^{m_1})^\top \leftarrow \mathrm{Vec}\left(\left(\left(x^{j\ell}\right)_{\substack{j=1,\dots,n_1 \\ \ell=1,\dots,\tau_1}}\right)^\top\right)$
  for $k = 2$ to $t$ do
    for $j = n_{k-1} + 1$ to $n_k$ do
      for $\ell = 1$ to $\tau_{k-1}$ do
        Sample $x^{j\ell}$ from $\chi_j$
      end for
    end for
    for $j = 1$ to $n_k$ do
      for $\ell = \tau_{k-1} + 1$ to $\tau_k$ do
        Sample $x^{j\ell}$ from $\chi_j$
      end for
    end for
    $(x^1, \dots, x^{m_k})^\top \leftarrow \mathrm{Vec}\left(\left(\left(x^{j\ell}\right)_{\substack{j=1,\dots,n_k \\ \ell=1,\dots,\tau_k}}\right)^\top\right)$
  end for
```

Algorithm 2. Random sequential sampling.

```
INPUT: $t$, $(\mu_{n_k})_{k=1}^{t-1}$, $(\sigma_{n_k})_{k=1}^{t}$
OUTPUT: $x^1, \dots, x^{m_t}$ s.t. $x^1, \dots, x^{m_k} \overset{\text{i.i.d.}}{\sim} \mu_{n_k}, k = 1, \dots, t$.
  for $j = 1$ to $m_1$ do
    Sample $x^j$ from $\mu_{n_1} = \sigma_{n_1}$
  end for
  for $k = 2$ to $t$ do
    Sample $B_k$ from Bin $\left(m_k, \dfrac{n_k - n_{k-1}}{n_k}\right)$
    for $j = \min(m_k - B_k, m_{k-1}) + 1$ to $\min(m_k - B_k, m_{k-1}) + \max(m_k - B_k - m_{k-1}, 0)$
    do
      Sample $x^j$ from $\mu_{n_{k-1}}$
    end for
    for $j = \min(m_k - B_k, m_{k-1}) + \max(m_k - B_k - m_{k-1}, 0) + 1$ to $m_k$ do
      Sample $x^j$ from $\sigma_{n_k}$
    end for
  end for
```

Algorithm 3. Adaptive WLS.

```
INPUT: $\Lambda_1 = \{(0, \dots, 0)^\top\}$, $\beta$, $s$, $\alpha$, $t$, $k_{\text{sg}}$
OUTPUT: $u_C^t$
  $\tau_1 = \lceil \theta^{-1} \ln(\zeta(s)(\#(\Lambda_1))^{s+1}/\alpha) \rceil$
  for each $\nu \in \Lambda_1$ do
    Add $\tau_1$ random samples distributed as $\chi_\nu$
  end for
  $m_1 = \tau_1 \#(\Lambda_1)$
  $u_C^1 = \text{argmin}_{v \in V_{\Lambda_1}} \|u - v\|_{m_1}$
  $r_1 = u - u_C^1$
  for $k = 2$ to $t$ do
    $F = \text{BULK}(\mathcal{R}(\Lambda_{k-1}), |\langle r_{k-1}, \psi_\nu \rangle_{m_{k-1}}|^2, \beta)$
    $\Lambda_k = \Lambda_{k-1} \cup F$
    $\tau_k = \lceil \theta^{-1} \ln(\zeta(s)(\#(\Lambda_k))^{s+1}/\alpha) \rceil$
    for each $\nu \in \Lambda_{k-1}$ do
      Add $\tau_k - \tau_{k-1}$ random samples distributed as $\chi_\nu$
    end for
    for each $\nu \in \Lambda_k \setminus \Lambda_{k-1}$ do
      Add $\tau_k$ random samples distributed as $\chi_\nu$
    end for
    $m_k = \tau_k \#(\Lambda_k)$
    $u_C^k = \text{argmin}_{v \in V_{\Lambda_k}} \|u - v\|_{m_k}$
    if $k$ mod $k_{sg} = 0$ then
      $\Lambda_k = \Lambda_{k-1} \cup \{\nu\}$, with $\nu$ being the most ancient multi-index in $\mathcal{R}(\Lambda_{k-1}) \setminus F$
    end if
    $r_k = u - u_C^k$
  end for
```

Algorithm 4. Fully adaptive WLS.

```
INPUT: $\Lambda_1 = \{(0,\dots,0)^\top\}$, $\beta$, $t$, $\xi$, $k_{\mathrm{sg}}$
OUTPUT: $u_C^t$
  repeat
    for each $\nu \in \Lambda_1$ do
      Add one random sample distributed as $\chi_\nu$
    end for
    $m_1 = m_1 + \#(\Lambda_1)$
  until $|||G_1 - I_1||| < \xi$
  $u_C^1 = \mathrm{argmin}_{v \in V_{\Lambda_1}} \|u - v\|_{m_1}$
  $r_1 = u - u_C^1$
  for $k = 2$ to $t$ do
    $F = \mathrm{BULK}(\mathcal{R}(\Lambda_{k-1}), |\langle r_{k-1}, \psi_\nu \rangle_{m_{k-1}}|^2, \beta)$
    $\Lambda_k = \Lambda_{k-1} \cup F$
    for each $\nu \in \Lambda_k \setminus \Lambda_{k-1}$ do
      Add $m_{k-1}/\#(\Lambda_{k-1})$ random samples distributed as $\chi_\nu$
    end for
    $m_k = m_{k-1}\#(\Lambda_k)/\#(\Lambda_{k-1})$
    repeat
      for each $\nu \in \Lambda_k$ do
        Add one random sample distributed as $\chi_\nu$
      end for
      $m_k = m_k + \#(\Lambda_k)$
    until $|||G_k - I_k||| < \xi$
    $u_C^k = \mathrm{argmin}_{v \in V_{\Lambda_k}} \|u - v\|_{m_k}$
    if $k$ mod $k_{sg} = 0$ then
      $\Lambda_k = \Lambda_{k-1} \cup \{\nu\}$, with $\nu$ being the most ancient multi-index in $\mathcal{R}(\Lambda_{k-1}) \setminus F$
    end if
    $r_k = u - u_C^k$
  end for
```

REFERENCES

[1] R. Ahlswede and A. Winter, *Strong converse for identification via quantum channels*, IEEE Trans. Inform. Theory, 48 (2002), pp. 569–579.

[2] B. Arras, M. Bachmayr, and A. Cohen, *Sequential Sampling for Optimal Weighted Least Squares Approximations in Hierarchical Spaces*, arXiv:1805.10801, 2018.

[3] H. Bungartz and M. Griebel, *Sparse grids*, Acta Numer., 13 (2004), pp. 147–269.

[4] G. Cybenko, *Approximations by superpositions of sigmoidal functions*, Math. Control Signals Systems, 2 (1989), pp. 303–314.

[5] A. Cohen, M. A. Davenport, and D. Leviatan, *On the stability and accuracy of least squares approximations*, Found. Comput. Math., 13 (2013), pp. 819–834.

[6] A. Cohen and R. DeVore, *Approximation of high-dimensional parametric PDEs*, Acta Numer., 24 (2015), pp. 1–159.

[7] A. Cohen and G. Migliorati, *Optimal weighted least-squares methods*, SMAI J. Comput. Math., 3 (2017), pp. 181–203.

[8] A. Cohen and G. Migliorati, *Multivariate Approximation in Downward Closed Polynomial Spaces*, in Contemporary Computational Mathematics—A Celebration of the 80th Birthday of Ian Sloan, Springer, New York, 2018.

[9] P. J. Davis, *Interpolation and Approximation*, Dover, Mineola, NY, 1975.

[10] R. DeVore and V. N. Temlyakov, *Some remarks on greedy algorithms*, Adv. Comput. Math., 5 (1996), pp. 173–187.

[11] A. Doostan and J. Hampton, *Coherence motivated sampling and convergence analysis of least squares polynomial Chaos regression*, Comput. Methods Appl. Mech. Engrg., 290 (2015), pp. 73–97.

[12] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Birkhäuser, Basel, 2013.

[13] J. D. Jakeman, A. Narayan, and T. Zhou, *A Christoffel function weighted least squares algorithm for collocation approximations*, Math. Comp., 86 (2017), pp. 1913–1947.

[14] T. Gerstner and M. Griebel, *Dimension-adaptive tensor-product quadrature*, Computing, 71 (2003), pp. 65–87.

[15] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer, New York, 2002.

[16] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function*, Neural Networks, 6 (1993), pp. 861–867.

[17] G. Migliorati, F. Nobile, E. von Schwerin, and R. Tempone, *Analysis of discrete L^2 projection on polynomial spaces with random evaluations*, Found. Comput. Math., 14 (2014), pp. 419–456.

[18] G. Migliorati, *Adaptive polynomial approximation by means of random discrete least squares*, Proceedings of ENUMATH 2013, Lect. Notes Comput. Sci. Eng. 103, Springer, New York, 2015, pp. 547–554.

[19] G. Migliorati, F. Nobile, and R. Tempone, *Convergence estimates in probability and in expectation for discrete least squares with noisy evaluations at random points*, J. Multivariate Anal., 142 (2015), pp. 167–182.

[20] J. Tropp, *User friendly tail bounds for sums of random matrices*, Found. Comput. Math., 12 (2012), pp. 389–434.