SIAM J. MATH. DATA SCI.
Vol. 1, No. 1, pp. 189–207

Sequential Sampling for Optimal Weighted Least Squares Approximations in Hierarchical Spaces*

Benjamin Arras†, Markus Bachmayr‡, and Albert Cohen§

Abstract. We consider the problem of approximating an unknown function $u \in L^2(D,\rho)$ from its evaluations at given sampling points $x^1,\dots,x^n \in D$, where $D \subset \mathbb{R}^d$ is a general domain and ρ a probability measure. The approximation is picked in a linear space V_m, where $m = \dim(V_m)$, and computed by a weighted least squares method. Recent results show the advantages of picking the sampling points at random according to a well-chosen probability measure μ that depends on both V_m and ρ. With such a random design, the weighted least squares approximation is proved to be stable with high probability, and having precision comparable to that of the exact $L^2(D,\rho)$-orthonormal projection onto V_m, in a near-linear sampling regime $n \sim m \log m$. The present paper is motivated by the adaptive approximation context, in which one typically generates a nested sequence of spaces $(V_m)_{m\geq 1}$ with increasing dimension. Although the measure $\mu = \mu_m$ changes with V_m, it is possible to recycle the previously generated samples by interpreting μ_m as a mixture between μ_{m-1} and an update measure σ_m. Based on this observation, we discuss sequential sampling algorithms that maintain the stability and approximation properties uniformly over all spaces V_m. Our main result is that the total number of computed samples at step m remains of the order $m \log m$ with high probability. Numerical experiments confirm this analysis.

Key words. weighted least squares, random matrices, optimal sampling measures, hierarchical approximation spaces, sequential sampling

AMS subject classifications. 41A10, 41A65, 62E17, 65C50, 93E24

DOI. 10.1137/18M1189749

1. Introduction.

Least squares approximations are ubiquitously used in numerical computation when trying to reconstruct an unknown function u defined on some domain $D \subseteq \mathbb{R}^d$ from its observations $y^1,\dots,y^n$ at a limited amount of points $x^1,\dots,x^n \in D$. In its simplest form the method amounts to minimizing the least squares fit

$$\frac{1}{n}\sum_{i=1}^{n} |y^i - v(x^i)|^2 \tag{1.1}$$

*Received by the editors May 29, 2018; accepted for publication (in revised form) October 15, 2018; published electronically February 12, 2019.

http://www.siam.org/journals/simods/1-1/M118974.html

Funding: The first author was supported by the European Research Council under grant ERC AdG 338977 BREAD. The second author was supported by the Hausdorff Center of Mathematics, University of Bonn. The third author was supported by the Institut Universitaire de France and by the European Research Council under grant ERC AdG 338977 BREAD.

†Laboratoire Paul Painlevé, Université de Lille, Cité Scientifique, 59655 Villeneuve d'Ascq, France (benjamin.arras@univ-lille.fr).

‡Institut für Mathematik, Johannes Gutenberg-Universität Mainz, 55128 Mainz, Germany (bachmayr@uni-mainz.de).

§Sorbonne Université, CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, 75005 Paris, France (cohen@ljll.math.upmc.fr).

over a set of functions v that are subject to certain constraints expressing a prior on the unknown function u. There are two classical approaches for imposing such constraints:

(i) Add a penalty term $P(v)$ to the least squares fit. Classical instances include norms of reproducing kernel Hilbert spaces or ℓ^1 norms that promote sparsity of v when expressed in a certain basis of functions.

(ii) Limit the search of v to a space V_m of finite dimension $m \leq n$. Classical instances include spaces of algebraic or trigonometric polynomials, wavelets, or splines.

The present paper is concerned with the second approach, in which approximability by the space V_m may be viewed as a prior on the unknown function. We measure accuracy in the Hilbertian norm

$$\|v\| = \left(\int_D |v(x)|^2 d\rho\right)^{1/2} = \|v\|_{L^2(D,\rho)}, \tag{1.2}$$

where ρ is a probability measure over D. We denote by $\langle \cdot, \cdot \rangle$ the associated inner product. The error of best approximation is defined by

$$e_m(u) := \min_{v \in V_m} \|u - v\| \tag{1.3}$$

and is attained by $P_m u$, the $L^2(D,\rho)$-orthogonal projection of u onto V_m. Since the least squares approximation $\tilde{u}$ is picked in V_m, it is natural to compare $\|u - \tilde{u}\|$ with $e_m(u)$. In particular, the method is said to be near-optimal (or instance optimal with constant C) if the comparison

$$\|u - \tilde{u}\| \leq C e_m(u) \tag{1.4}$$

holds for all u, where $C > 1$ is some fixed constant.

The present paper is motivated by applications where the sampling points x^i are not prescribed and can be chosen by the user. Such a situation typically occurs when evaluation of u is performed by either computer simulation or physical experiment, depending on a vector of input parameters x that can be set by the user. This evaluation is typically costly, and the goal is to obtain a satisfactory *surrogate model* $\tilde{u}$ from a minimal number of evaluations

$$y^i = u(x^i). \tag{1.5}$$

For a given probability measure ρ and approximation space V_m of interest, a relevant question is therefore whether instance optimality can be achieved with sample size n that is moderate, ideally linear in m.

Recent results of [11, 12, 17] for polynomial spaces and [6] in a general approximation setting show that this objective can be achieved by certain random sampling schemes in the more general framework of *weighted least squares* methods. The approximation $\tilde{u}$ is then defined as the solution to

$$\min_{v \in V_m} \frac{1}{n} \sum_{i=1}^{n} w(x^i) |y^i - v(x^i)|^2, \tag{1.6}$$

where w is a positive function and the x^i are independently drawn according to a probability measure μ, satisfying the constraint

$$w\, d\mu = d\rho. \tag{1.7}$$

The choice of a sampling measure μ that differs from the error norm measure ρ appears to be critical in order to obtain instance optimal approximations with an optimal sampling budget.

In particular, it is shown in [12, 6] that there exists an optimal choice of (ρ, μ) such that the weighted least squares is stable with high probability and instance optimal in expectation, under the near-linear regime $n \sim m$ up to logarithmic factors. The optimal sampling measure and weights are given by

$$d\mu_m = \frac{k_m}{m} d\rho \quad \text{and} \quad w_m = \frac{m}{k_m}, \tag{1.8}$$

where k_m is the so-called Christoffel function defined by

$$k_m(x) = \sum_{j=1}^{m} |\varphi_j(x)|^2, \tag{1.9}$$

with $\{\varphi_1, \dots, \varphi_m\}$ any $L^2(D, \rho)$-orthonormal basis of V_m.

In many practical applications, the space V_m is picked within a family $(V_m)_{m \geq 1}$ that has the nestedness property

$$V_1 \subset V_2 \subset \cdots, \tag{1.10}$$

and accuracy is improved by raising the dimension m. The basis $(\varphi_1, \dots, \varphi_m)$ is therefore the section corresponding to the m first elements within an infinite orthonormal sequence $(\varphi_k)_{k \geq 1}$.

The sequence $(V_m)_{m \geq 1}$ may be either a priori defined or adaptively generated, which means that the way V_m is refined into V_{m+1} may depend on the result of the least squares computation. Examples of such hierarchical adaptive or nonadaptive schemes include in particular the following:

(i) Mesh refinement in low-dimension performed by progressive addition of hierarchical basis functions or wavelets, which is relevant for approximating piecewise smooth functions, such as images or shock profiles; see [3, 9, 10].

(ii) Sparse polynomial approximation in high dimension, which is relevant for the treatment of certain parametric and stochastic PDEs; see [2, 4, 5, 7].

Another typical setting is when we are given an orthonormal basis $(\psi_j)_{j \geq 1}$ of $L^2(D, \rho)$ and want to approximate u by a combination of m functions from this basis. In a nonadaptive setting, the spaces V_m are simply defined as the span of the m first elements; that is, we take $\varphi_j = \psi_j$. In the adaptive setting, the spaces V_m are spanned by m arbitrary elements $\psi_{j_1}, \dots, \psi_{j_m}$, so that $\varphi_k = \psi_{j_k}$, and the next choice of j_{m+1} depends on the result of the least squares computation.

In all such settings, we are faced with the difficulty that the optimal measure μ_m defined by (1.8) varies together with m.

In order to maintain an optimal sampling budget, one should avoid the option of drawing a new sample

$$S_m = \{x_m^1, \dots, x_m^n\} \tag{1.11}$$

of increasing size $n = n(m)$ at each step m. In the particular case where V_m are the univariate polynomials of degree $m-1$, and ρ is a Jacobi-type measure on $[-1,1]$, it is known that μ_m converges weakly to the equilibrium measure defined by

$$d\mu^*(y) = \frac{dy}{\pi\sqrt{1-y^2}}, \tag{1.12}$$

with a uniform upper bound

$$\mu_m \le c\mu^*; \tag{1.13}$$

see [20, 18]. This suggests the option of replacing all μ_m by the single μ^*, as studied in [17]. Unfortunately, such an asymptotic behavior is not encountered for most general choices of spaces $(V_m)_{m\ge 1}$. An estimate of the form (1.13) was proved in [15] for sparse multivariate polynomials, however, with a c that increases exponentially with the dimension, which theoretically impacts in a similar manner the sampling budget needed for stability.

In this paper, we discuss sampling strategies that are based on the observation that the optimal measure μ_m enjoys the mixture property

$$\mu_{m+1} = \Big(1 - \frac{1}{m+1}\Big)\mu_m + \frac{1}{m+1}\sigma_{m+1}, \quad \text{where} \quad d\sigma_m := |\varphi_m|^2 d\rho, \tag{1.14}$$

which readily follows from (1.8) and (1.9). As noticed in [16], this leads naturally to sequential sampling strategies, where the sample S_m is recycled for generating S_{m+1}. The main contribution of this paper is to analyze such a sampling strategy and prove that the two following properties can be jointly achieved in an expectation or high-probability sense:

1. Stability and instance optimality of weighted least squares hold uniformly over all $m \ge 1$.
2. The total sampling budget after m steps is linear in m up to logarithmic factors.

The rest of the paper is organized as follows. We recall in section 2 stability and approximation estimates from [12, 6] concerning the weighted least squares method in a fixed space V_m. We then describe in section 3 a sequential sampling strategy based on (1.14) and establish the above optimality properties 1 and 2. These optimality properties are numerically illustrated in section 4 for the algorithm and two of its variants.

2. Optimal weighted least squares.

We denote by $\|\cdot\|_n$ the discrete Euclidean norm defined by

$$\|v\|_n^2 := \frac{1}{n}\sum_{i=1}^n w(x^i)|v(x^i)|^2, \tag{2.1}$$

and by $\langle \cdot, \cdot \rangle_n$ the associated inner product. The solution $\tilde{u} \in V_m$ to (1.6) may be thought of as an orthogonal projection of u onto V_m for this norm. Expanding

$$\tilde{u} = \sum_{j=1}^{m} c_j \varphi_j \tag{2.2}$$

in the basis $\{\varphi_1, \dots, \varphi_m\}$ of V_m, the coefficient vector $\mathbf{c} = (c_1, \dots, c_m)^T$ is a solution to the linear system

$$\mathbf{G}_m \mathbf{c} = \mathbf{d}, \tag{2.3}$$

where $\mathbf{G}_m$ is the Gramian matrix for the inner product $\langle \cdot, \cdot \rangle_n$ with entries

$$\mathbf{G}_{j,k} := \langle \varphi_j, \varphi_k \rangle_n = \frac{1}{n} \sum_{i=1}^{n} w(x^i) \varphi_j(x^i) \varphi_k(x^i), \tag{2.4}$$

and the vector $\mathbf{d}$ has entries $\mathbf{d}_k = \frac{1}{n} \sum_{i=1}^{n} w(x^i) y^i \varphi_k(x^i)$. The solution $\mathbf{c}$ always exists and is unique when $\mathbf{G}_m$ is invertible.

Since $x^1, \dots, x^n$ are drawn independently according to $d\mu$, the relation (1.7) implies that $\mathbb{E}(\langle v_1, v_2 \rangle_n) = \langle v_1, v_2 \rangle$, and in particular

$$\mathbb{E}(\mathbf{G}_m) = \mathbf{I}. \tag{2.5}$$

The stability and accuracy analysis of the weighted least squares method can be related to the amount of deviation between $\mathbf{G}_m$ and its expectation $\mathbf{I}$ measured in the spectral norm. Recall that for $m \times m$ matrices $\mathbf{M}$, this norm is defined as $\|\mathbf{M}\|_2 = \sup_{\|v\|_2 = 1} \|\mathbf{M}v\|_2$. This deviation also describes the closeness of the norms $\|\cdot\|$ and $\|\cdot\|_n$ over the space V_m, since one has

$$\|\mathbf{G}_m - \mathbf{I}\|_2 \leq \delta \quad \Longleftrightarrow \quad (1-\delta)\|v\|^2 \leq \|v\|_n^2 \leq (1+\delta)\|v\|^2, \quad v \in V_m. \tag{2.6}$$

Note that this closeness also implies a bound

$$\kappa(\mathbf{G}_m) \leq \frac{1+\delta}{1-\delta} \tag{2.7}$$

on the condition number of $\mathbf{G}_m$.

Following [6], we use the particular value $\delta = \frac{1}{2}$ and define $\tilde{u}$ as the solution to (1.6) when $\|\mathbf{G}_m - \mathbf{I}\|_2 \leq \frac{1}{2}$ and set $\tilde{u} = 0$ otherwise. The probability of the latter event can be estimated by a matrix tail bound, noting that

$$\mathbf{G}_m = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}^i, \tag{2.8}$$

where the $\mathbf{X}^i$ are n independent realizations of the rank-one random matrix

$$\mathbf{X} := w(x)(\varphi_j(x)\varphi_k(x))_{j,k=1,\dots,m}, \tag{2.9}$$

where x is distributed according to μ.

The matrix Chernoff bound (see [21, Theorem 1.1]) says that if K is an almost sure bound for the spectral norm $\|\mathbf{X}\|_2$, then, for any $0 < \delta < 1$,

$$\Pr\{\|\mathbf{G}_m - \mathbf{I}\|_2 > \delta\} \leq 2m \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{1/K} = 2m \exp\left(-\frac{c_\delta}{K}\right), \tag{2.10}$$

with $c_\delta := (1+\delta)\ln(1+\delta) - \delta > 0$. Taking $\delta = \frac{1}{2}$ and observing that $K := \|wk_m\|_{L^\infty}$ is an almost sure bound for $\|\mathbf{X}\|_2$, we obtain

$$\Pr\Big(\|\mathbf{G}_m - \mathbf{I}\|_2 \geq \frac{1}{2}\Big) \leq 2m \exp(-\gamma n/K), \quad \gamma := \frac{1}{2}\big(3\ln(3/2) - 1\big). \tag{2.11}$$

Since $\int wk_m d\mu = \int k_m d\rho = m$, it follows that K is always larger than m. With the choice $\mu = \mu_m$ given by (1.8) for the sampling measure, one has exactly $K = m$, which leads to the following result.

Theorem 2.1. *Assume that the sampling measure and weight function are given by* (1.8). *Then for any* $0 < \varepsilon < 1$, *the condition*

$$n \geq cm(\ln(2m) - \ln(\varepsilon)), \quad c := \gamma^{-1} = \frac{2}{3\ln(3/2) - 1}, \tag{2.12}$$

implies the following stability and instance optimality properties:

$$\Pr\Big(\|\mathbf{G}_m - \mathbf{I}\|_2 \geq \frac{1}{2}\Big) \leq \varepsilon, \quad m \geq 1. \tag{2.13}$$

In addition, one has the instance optimality property

$$\mathbb{E}(\|u - \tilde{u}\|^2) \leq \Big(1 + \frac{c}{\ln(2m) - \ln(\varepsilon)}\Big) e_m(u)^2 + \varepsilon\|u\|^2. \tag{2.14}$$

The detailed proof of the instance optimality property (2.14) may be found in [6]. Here we provide a simple argument that leads to a similar estimate with a less sharp multiplicative constant: in the event where $\|\mathbf{G}_m - \mathbf{I}\|_2 \leq \frac{1}{2}$, one has

$$\begin{aligned} \|u - \tilde{u}\|^2 &= \|u - P_m u\|^2 + \|\tilde{u} - P_m u\|^2 \\ &\leq e_m(u)^2 + 2\|\tilde{u} - P_m u\|_n^2 \leq e_m(u)^2 + 2\|u - P_m u\|_n^2, \end{aligned} \tag{2.15}$$

and therefore

$$\mathbb{E}(\|u - \tilde{u}\|^2) \leq e_m(u)^2 + 2\mathbb{E}(\|u - P_m u\|_n^2) + \varepsilon\|u\|^2 = 3e_m(u)^2 + \varepsilon\|u\|^2. \tag{2.16}$$

Remark 2.2. One should make a clear distinction between the above instance optimality results and the classical results on nonparametric least squares bounded regression, such as those from [8] or in [13, Chapter 11]. Indeed, the latter consider the setting where the distribution ρ of the x^i is imposed and unknown to us. The convergence estimates towards the regression function u are stated in the norm $L^2(D, \rho)$, but do not provide instance optimality. In the present work, we operate in a different setting, sometimes called active learning, since we allow ourselves to choose the sampling strategy. As explained, the choice of the optimal sampling measure is critical in the derivation of our instance optimality estimates.

In summary, when using the optimal sampling measure μ_m defined by (1.8), stability and instance optimality can be achieved in the near-linear regime

$$n = n_\varepsilon(m) := \lceil c\, m\, (\ln(2m) - \ln \varepsilon) \rceil, \tag{2.17}$$

where $\varepsilon \in]0,1[$ controls the probability of failure. To simplify notation later, we set $n_\varepsilon(0) := 0$.

3. An optimal sequential sampling procedure. In the following analysis of a sequential sampling scheme, we assume a sequence of nested spaces $V_1 \subset V_2 \subset \cdots$ and corresponding basis functions $\varphi_1, \varphi_2, \ldots$ to be given such that for each m, $\{\varphi_1, \ldots, \varphi_m\}$ is an orthonormal basis of V_m. In practical applications, such spaces may either be fixed in advance or adaptively selected, that is, the choice of φ_{m+1} depends on the computation of the weighted least squares approximation (1.6) for V_m. In view of the previous result, one natural objective is to generate sequences of samples $(S_m)_{m\geq 1}$ distributed according to the different measures $(\mu_m)_{m\geq 1}$, with

$$\#(S_m) = n_\varepsilon(m), \tag{3.1}$$

for some prescribed $\varepsilon > 0$.

The simplest option for generating such sequences would be directly drawing samples from μ_m for each $m = 1, 2, \ldots$ separately. Since we ask that $n_\varepsilon(m)$ is proportional to m up to logarithmic factors, this leads to a total cost C_m after m step given by

$$C_m = \sum_{k=1}^{m} \#(S_k), \tag{3.2}$$

which increases faster than quadratically with m. Instead, we want to recycle the existing samples S_m in generating S_{m+1} to arrive at a scheme such that the total cost C_m remains comparable to $n_\varepsilon(m)$, that is, close to linear in m.

To this end, we use the mixture property (1.14). In what follows, we assume a procedure for sampling from each update measure $d\sigma_j := |\varphi_j|^2 d\rho$ to be available. In the univariate case, standard methods are inversion transform sampling or rejection sampling. These methods may in turn serve in the multivariate case when the φ_j are tensor product basis functions on a product domain and ρ is itself of tensor product type, since σ_j are then product measures that can be sampled via their univariate factor measures. We first observe that, in order to draw x distributed according to μ_m, for some fixed m, we can proceed as follows:

(3.3) Draw j uniformly distributed in $\{1, \ldots, m\}$, then draw x from σ_j.

Now let $(n(m))_{m\geq 1}$ be an increasing sequence representing the prescribed size of the samples $(S_m)_{m\geq 1}$. Suppose that we are given $S_m = \{x_m^1, \ldots, x_m^{n(m)}\}$ i.i.d. according to μ_m. In order to obtain the new sample $S_{m+1} = \{x_{m+1}^1, \ldots, x_{m+1}^{n(m+1)}\}$ i.i.d. according to μ_{m+1}, we can proceed as stated in Algorithm 1.

Algorithm 1 requires a fixed number $n(m+1) - n(m)$ of samples from μ_{m+1} and an

Algorithm 1. Sequential sampling.

input: sample $S_m = \{x_m^1, \dots, x_m^{n(m)}\}$ from μ_m
output: sample $S_{m+1} = \{x_{m+1}^1, \dots, x_{m+1}^{n(m+1)}\}$ from μ_{m+1}

```
for i = 1, ..., n(m) do
    draw a_i uniformly distributed in {1, ..., m + 1}
    if a_i = m + 1 then
        draw x^i_{m+1} from σ_{m+1}
    else
        set x^i_{m+1} := x^i_m
    end if
end for
for i = n(m) + 1, ..., n(m + 1) do
    draw x^i_{m+1} from μ_{m+1} by (3.3)
end for
```

additional number $\tilde{n}(m)$ of samples from σ_{m+1}. The latter is a random variable that can be expressed as

$$\tilde{n}(m) = \sum_{i=1}^{n(m)} b_{m+1}^i, \tag{3.4}$$

where for each fixed $m \geq 1$, the $(b_m^i)_{i=1,\dots,n(m)}$ are i.i.d. Bernoulli random variables with $\Pr(b_m^i = 1) = \frac{1}{m}$. Moreover, $\{b_m^i : i = 1, \dots, n(m),\ m \geq 1\}$ is a collection of independent random variables. This immediately gives an expression for the total cost C_m after m successive applications of Algorithm 1, beginning with $n(1)$ samples from μ_1, as the random variable

$$C_m := n(m) + s(m), \quad s(m) := \sum_{k=1}^{m-1} \tilde{n}(k) = \sum_{k=1}^{m-1} \sum_{i=1}^{n(k)} b_{k+1}^i. \tag{3.5}$$

We now focus on the particular choice

$$n(m) := n_\varepsilon(m), \tag{3.6}$$

as in (2.17), for a prescribed $\varepsilon \in]0, 1[$. This particular choice ensures that, for all $m \geq 1$,

$$\Pr\big(\|\mathbf{G}_m - \mathbf{I}\|_2 \geq \tfrac{1}{2}\big) \leq \varepsilon, \tag{3.7}$$

where $\mathbf{G}_m$ denotes the Gramian for V_m according to (2.4).

We first estimate $\mathbb{E}(s(m))$ for this choice of $n(m)$. For this purpose, we note that $\mathbb{E}(\tilde{n}(k)) = \frac{n(k)}{k+1}$ and use the following lemma.

Lemma 3.1. *For $m \geq 1$ and $\varepsilon > 0$,*

$$\tfrac{1}{2} n(m) - 2c \leq \sum_{k=1}^{m} \frac{n(k)}{k+1} \leq n(m) + 1. \tag{3.8}$$

Proof. For the upper bound, we note that $n(k) \le ck(\ln(2k) - \ln\varepsilon) + 1$ and

$$\sum_{k=1}^{m} \frac{1}{k+1}\big(ck(\ln(2k) - \ln\varepsilon) + 1\big) \le c\sum_{k=1}^{m}\big(\ln(2k) - \ln\varepsilon\big) + \sum_{k=1}^{m}\frac{1}{k+1} \le c\big(m\ln 2 + \ln m! - m\ln\varepsilon\big) + \ln(m+1). \tag{3.9}$$

By the Stirling bound $k! \le ek^{k+\frac{1}{2}}e^{-k}$ for $k \ge 1$,

$$\ln m! \le m\big(\ln m - 1\big) + \tfrac{1}{2}\ln m + 1. \tag{3.10}$$

Combining this with (3.9) gives

$$\sum_{k=1}^{m}\frac{n(k)}{k+1} \le cm\big(\ln 2m - \ln\varepsilon\big) - c(m-1) + \frac{c}{2}\ln m + \ln(m+1) \le n(m) + 1, \tag{3.11}$$

where the inequality $-c(m-1) + \frac{c}{2}\ln m + \ln(m+1) \le 1$ for $m \ge 1$ is verified for the choice of c in (2.12) by direct evaluation for $m = 1$ and monotonicity. For the lower bound, we estimate

$$\sum_{k=1}^{m}\frac{n(k)}{k+1} \ge \frac{cm}{2}(\ln 2 - \ln\varepsilon) + c\sum_{k=1}^{m}\frac{k}{k+1}\ln k. \tag{3.12}$$

Using monotonicity and integration by parts with $\frac{d}{dx}\,(x - \ln(1+x)) = \frac{x}{x+1}$, we obtain

$$\begin{aligned}\sum_{k=1}^{m}\frac{k}{k+1}\ln k &\ge \int_1^m \frac{x}{x+1}\ln x\,dx \\ &= -\int_1^m \frac{x - \ln(x+1)}{x}\,dx + \big[(x - \ln(x+1))\ln x\big]_1^m \\ &= -(m-1) + \int_1^m \frac{\ln(x+1)}{x}\,dx + (m - \ln(m+1))\ln m.\end{aligned} \tag{3.13}$$

Moreover,

$$\int_1^m \frac{\ln(x+1)}{x}\,dx = \int_0^{\ln m}\ln(1+e^t)\,dt \ge \int_0^{\ln m} t\,dt = \frac{1}{2}\ln^2 m, \tag{3.14}$$

and using this in (3.13) gives

$$\sum_{k=1}^{m}\frac{k}{k+1}\ln k \ge \frac{1}{2}m\ln m + \frac{1}{2c} + R(m), \tag{3.15}$$

where

$$R(m) = \frac{1}{2}m(\ln m - 2) - \ln(m+1)\ln m + \frac{1}{2}\ln^2 m + 1 - \frac{1}{2c} > -2. \tag{3.16}$$

The latter inequality can be directly verified for the first few values of m and then follows for $m \geq 1$ by monotonicity. Thus, we obtain

$$\sum_{k=1}^{m} \frac{n(k)}{k+1} \geq \tfrac{1}{2}\big[cm\big(\ln 2m - \ln \varepsilon\big) + 1\big] - 2c \geq \tfrac{1}{2} n(m) - 2c, \tag{3.17}$$

which concludes the proof of the lemma. ∎

As an immediate consequence, we obtain an estimate of the total cost in expectation by the following.

Theorem 3.2. *With $n(m) = n_\varepsilon(m)$, the total cost after m steps of Algorithm 1 satisfies*

$$n(m) + \tfrac{1}{2} n(m-1) - 2c \leq \mathbb{E}(C_m) \leq n(m) + n(m-1) + 1. \tag{3.18}$$

We next derive estimates for the probability that the upper bound in the above estimate is exceeded substantially by C_m. The following Chernoff inequality can be found in [1]. For the reader's convenience, we give a standard short proof following [14].

Lemma 3.3. *Let $N \geq 1$, $(p_i)_{1 \leq i \leq N} \in [0,1]^N$, and $(X_i)_{1 \leq i \leq N}$ be a collection of independent Bernoulli random variables with $\Pr(X_i = 1) = p_i$ for all $1 \leq i \leq N$. Set $\bar{X} = \mathbb{E}(X_1 + \cdots + X_N) = p_1 + \cdots + p_N$. Then, for all $\tau \geq 0$,*

$$\Pr\big(X_1 + \cdots + X_N \geq (1+\tau)\bar{X}\big) \leq (1+\tau)^{-(1+\tau)\bar{X}} e^{\tau \bar{X}}.$$

In particular, for all $\tau \in [0,1]$,

$$\Pr\big(X_1 + \cdots + X_N \geq (1+\tau)\bar{X}\big) \leq e^{-\tau^2 \bar{X}/3}.$$

Proof. For $t \geq 0$ and $Y := X_1 + \cdots + X_N$, $\Pr(Y \geq (1+\tau)\bar{X}) \leq e^{-t(1+\tau)\bar{X}} \mathbb{E}(e^{tY})$ by Markov's inequality, and using that $1 + a \leq e^a$, $a \in \mathbb{R}$,

$$\mathbb{E}(e^{tY}) = \prod_{i=1}^{N} \mathbb{E}(e^{tX_i}) = \prod_{i=1}^{N} \big(p_i e^t + (1-p_i)\big) \leq \prod_{i=1}^{N} e^{p_i(e^t-1)} = e^{(e^t-1)\bar{X}}.$$

Now take $t = \ln(1+\tau)$. Moreover, for $\tau \in [0,1]$, one has $\tau - (1+\tau)\ln(1+\tau) \leq -\frac{1}{3}\tau^2$. ∎

We now apply this result to the random part $s(m)$ of the total sampling costs C_m as defined in (3.5) and obtain the following probabilistic estimate.

Theorem 3.4. *With $n(m) = n_\varepsilon(m)$, the total cost after m steps of Algorithm 1 satisfies, for any $\tau \in [0,1]$,*

$$\begin{aligned} \Pr\big(C_m \geq n(m) + (1+\tau)(n(m-1)+1)\big) &\leq M_\tau e^{-\frac{\tau^2}{6} n(m-1)} \\ &\leq M_\tau \left(\frac{2(m-1)}{\varepsilon}\right)^{-\frac{\tau^2 c}{6}(m-1)} \end{aligned} \tag{3.19}$$

with $M_\tau := e^{\frac{2c\tau^2}{3}}$.

Proof. Applying Proposition 3.3 to the independent variable b_k^i which appears in $s(m)$, we obtain

$$\Pr\big(C_m \geq n(m) + (1+\tau)\mathbb{E}(s(m))\big) = \Pr\big(s(m) \geq (1+\tau)\mathbb{E}(s(m))\big) \leq e^{-\tau^2\mathbb{E}(s(m))/3}. \tag{3.20}$$

The result follows by using the lower bound on $\mathbb{E}(s(m))$ in Lemma 3.1. ∎

With the choice $n(m) = n_\varepsilon(m)$ we ensure that, for each value of m separately, the failure probability is bounded by ε. When the intermediate results in each step are used to drive an adaptive selection of the sequence of basis functions, however, it will typically be of interest to ensure the stronger uniform statement that, with high probability, $\|\mathbf{G}_m - \mathbf{I}\|_2 \leq \frac{1}{2}$ for *all* m. To achieve this, we now consider a slight modification of the above results with m-dependent choice of failure probability to ensure that $\mathbf{G}_m$ remains well-conditioned with high probability jointly for all m. We define the sequence of failure probabilities

$$\varepsilon(m) = \frac{6\varepsilon_0}{(\pi m)^2}, \quad m \geq 1, \tag{3.21}$$

for a fixed $\varepsilon_0 \in]0,1[$, and now analyze the repeated application of Algorithm 1, using

$$n(m) := n_{\varepsilon(m)}(m) = \lceil c\, m\, (\ln(2m) + 2\ln m - \ln(6\varepsilon_0/\pi^2)) \rceil \tag{3.22}$$

samples for V_m. Note that $n(m)$ differs only by a further term of order $\log m$ from $n_{\varepsilon_0}(m)$.

Lemma 3.5. *For $m \geq 1$, let $n(m)$ be defined as in (3.22) with $\varepsilon(m)$ as in (3.21). Then*

$$\tfrac{1}{2}n(m) - 6c \leq \sum_{k=1}^{m} \frac{n(k)}{k+1} \leq n(m) + 1. \tag{3.23}$$

Proof. For the upper bound in (3.23), note that $n(k) \leq ck\big(\ln(2k) + \ln k^2 - \ln(6\varepsilon_0/\pi^2)\big) + 1$. Using (3.10),

$$\sum_{k=1}^{m} \ln k^2 \leq m \ln m^2 + 2\big(1 - m + \tfrac{1}{2}\ln m\big) \leq m \ln m^2. \tag{3.24}$$

Thus, for all $m \geq 1$,

$$\begin{aligned}\sum_{k=1}^{m} \frac{n(k)}{k+1} &\leq c\sum_{k=1}^{m} \frac{k}{k+1}\left(\ln(2k) - \ln\left(\frac{6\varepsilon_0}{\pi^2}\right)\right) + \sum_{k=1}^{m}\frac{1}{k+1} + c\sum_{k=1}^{m}\frac{k\ln k^2}{1+k} \\ &\leq cm\big(\ln 2m - \ln(6\varepsilon_0/\pi^2)\big) + 1 + cm\ln m^2 \\ &\leq n(m) + 1,\end{aligned}$$

where we have used (3.9), (3.11) together with (3.24). Let us deal with the lower bound. For all $m \geq 1$,

$$\begin{aligned}\sum_{k=1}^{m} \frac{n(m)}{k+1} &\geq \sum_{k=1}^{m} \frac{1}{k+1}\left(ck\left(\ln(2k) - \ln\left(\frac{6\varepsilon_0}{\pi^2}\right) + \ln k^2\right)\right) \\ &\geq \frac{1}{2}\left(cm\left(\ln(2m) - \ln\left(\frac{6\varepsilon_0}{\pi^2}\right)\right) + 1\right) - 2c + \sum_{k=1}^{m} c\frac{k}{k+1}\ln k^2. \end{aligned} \tag{3.25}$$

Moreover using (3.15) and (3.16)

$$c\sum_{k=1}^{m}\frac{k}{k+1}\ln k^2 \geq 2c\left(\frac{m}{2}\ln m + \frac{1}{2c} - 2\right).$$

Thus, combining the previous bound together with (3.25) and the definition of $n(m)$ concludes the proof of the lemma. ∎

As a consequence, analogously to (3.18) we have

$$\mathbb{E}(C_m) \leq n(m) + n(m-1) + 1. \tag{3.26}$$

Using the above lemma as in Theorem 3.4 combined with a union bound, we arrive at the following uniform stability result.

Theorem 3.6. *Let $\varepsilon(m)$ be defined as in (3.21). Then, applying Algorithm 1 with $n(m)$ as in (3.22), one has*

$$\Pr\big(\exists m \in \mathbb{N}\colon \|\mathbf{G}_m - \mathbf{I}\|_2 \geq \tfrac{1}{2}\big) \leq \varepsilon_0,$$

and for any $\tau \in [0,1]$ and all $m \geq 1$, the random variable C_m satisfies

$$\begin{aligned}\Pr\big(C_m \geq n(m) + (1+\tau)(n(m-1)+1)\big) &\leq M_\tau e^{-\frac{\tau^2}{6}n(m-1)} \\ &\leq M_\tau\left(\frac{\pi^2(m-1)}{3\varepsilon_0}\right)^{-\frac{\tau^2 c}{2}(m-1)}\end{aligned} \tag{3.27}$$

with $M_\tau := e^{2c\tau^2}$.

In summary, applying Algorithm 1 successively to generate the samples $S_1, S_2, \ldots$, we can ensure that $\|\mathbf{G}_m - \mathbf{I}\|_2 \leq \frac{1}{2}$ holds uniformly for all steps with probability at least $1-\varepsilon_0$. The corresponding total costs for generating $S_1, \ldots, S_m$ can exceed a fixed multiple of $n(m)$ only with a probability that rapidly approaches zero as m increases.

Remark 3.7. Algorithm 1 and the above analysis can be adapted to create samples from μ_{m+q}, $q \geq 2$, using those for μ_m, which corresponds to adding q basis functions in one step. In this case,

$$\mu_{m+q} = \frac{m}{m+q}\mu_m + \frac{q}{m+q}\sum_{j=m+1}^{m+q}\frac{1}{q}\sigma_j, \tag{3.28}$$

where samples from $\sum_{j=m+1}^{m+q}\frac{1}{q}\sigma_j$ can be obtained by mixture sampling as in (3.3).

Remark 3.8. Another sequential sampling strategy was recently proposed in [19]. In this approach the $n = n(m)$ independent samples according to the optimal measure are replaced by m blocks of size $\lceil\frac{n(m)}{m}\rceil$. Each block j contains independent samples according to the measure σ_j, therefore amounting to a deterministic mixture. This strategy also allows us to perform sequential sampling while maintaining the optimal budget.

4. Numerical illustration. In our numerical tests, we consider two different types of orthonormal bases and corresponding target measures ρ on $D \subseteq \mathbb{R}$:

(i) On the one hand, we consider the case where D is equal to $\mathbb{R}$, ρ is the standard Gaussian measure on $\mathbb{R}$, and V_m is the vector space spanned by the Hermite polynomials normalized to one with respect to the norm $\|.\|$ up to degree $m-1$ for all $m \geq 1$. This case is an instance of the polynomial approximation method where the function u and its approximants from V_m might be unbounded in L^∞, as well as the Christoffel function $k_m(x) := \sum_{j=1}^m |H_{j-1}(x)|^2$.

(ii) On the other hand, we consider the case where $D = [0,1]$, with ρ the uniform measure on $[0,1]$, and the approximation spaces $(V_m)_{m\geq 1}$ are generated by Haar wavelet refinement. The Haar wavelets are of the form $\psi_{l,k} = 2^{l/2}\psi(2^l - k)$ with $l \geq 0$ and $k = 0, \ldots, 2^l - 1$, and $\psi := \chi_{[0,1/2[} - \chi_{[1/2,1[}$. In adaptive approximation, the spaces V_m are typically generated by including, as the scale level l grows, the values of k such that the coefficient of $\psi_{l,k}$ for the approximated function is expected to be large. This can be described by growing a finite tree within the hierarchical structure induced by the dyadic indexing of the Haar wavelet family: the indices $(l+1, 2k)$ or $(l+1, 2k+1)$ can be selected only if (l,k) has already been. In our experiment, we generate the spaces V_m by letting such a tree grow at random, starting from $V_1 = \text{span}\{\psi_{0,0}\}$. This selection of $(V_m)_{m\geq 1}$ is done once and used for all further tests. The resulting sampling measures $(\mu_m)_{m\geq 1}$ exhibit the local refinement behavior of the corresponding approximation spaces $(V_m)_{m\geq 1}$.

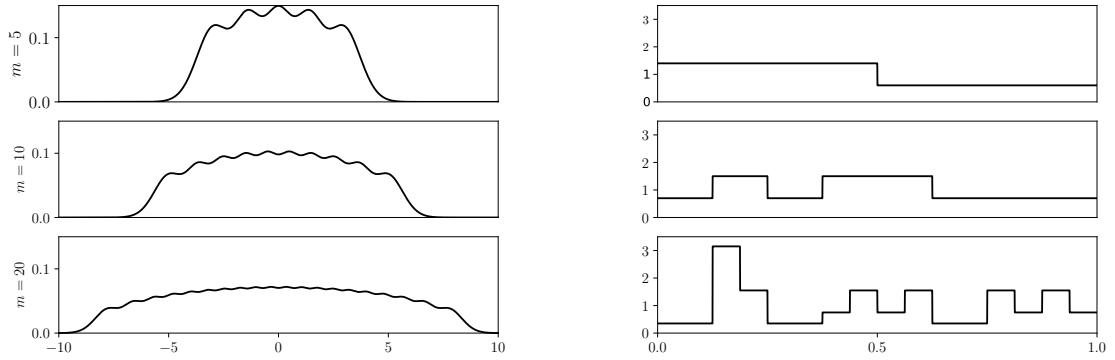
As seen further, although the spaces and measures are quite different, the cost of the sampling algorithm behaves similarly in these two cases. While we have used univariate domains D for the sake of numerical simplicity, we expect a similar behavior in multivariate cases, since our results are also immune to the spatial dimension d. As already mentioned, the sampling method is then facilitated when the functions φ_j are tensor product functions and ρ is a product measure.

The sampling measures μ_m are shown in Figure 1. In case (ii), the measures σ_j are uniform measures on dyadic intervals in $]0,1[$, from which we can sample directly, using $\mathcal{O}(1)$ operations per sample. In case (i), we have instead $d\sigma_j = |H_{j-1}|^2 d\rho$. Several strategies for sampling from these densities are discussed in [6]. In our following tests, we use inverse transform sampling: we draw z uniformly distributed in $[0,1]$ and obtain a sample x from σ_j by solving $\Phi_j(x) := \int_{-\infty}^x |H_{j-1}|^2 d\rho = z$. To solve these equations, in order to ensure robustness, we use a simple bisection scheme. Each point value of Φ_j can be obtained using $\mathcal{O}(j)$ operations, exploiting the fact that for $j \geq 2$,

$$\Phi_j = \Phi - \sum_{k=1}^{j-1} \frac{H_k H_{k-1}}{\sqrt{k}} g,$$

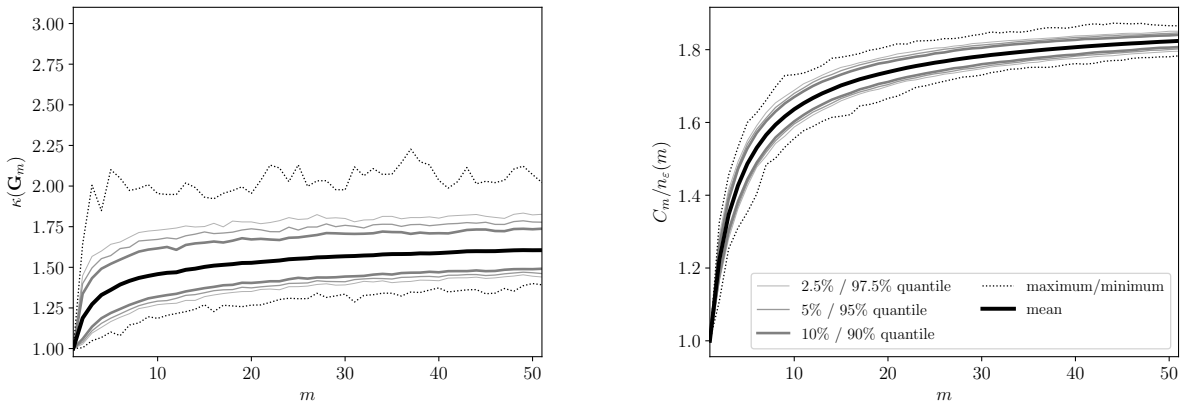
where g is the standard Gaussian density and Φ its cumulative distribution function. The bisection thus takes $\mathcal{O}(j \log \epsilon)$ operations to converge to accuracy ϵ. For practical purposes, the sampling for case (i) can be accelerated by precomputation of interpolants for σ_j.

In each of our numerical tests, we consider the distributions of $\kappa(\mathbf{G}_m)$ and C_m resulting from the sequential generation of sample sets S_m for $m = 1, \ldots, 50$ by Algorithm 1. In



(a) Hermite polynomials (b) Haar wavelets

Figure 1. *Sampling densities μ_m for* (a) *Hermite polynomials of degrees* $0, \dots, m-1$, (b) *subsets of Haar wavelet basis selected by random tree refinement.*

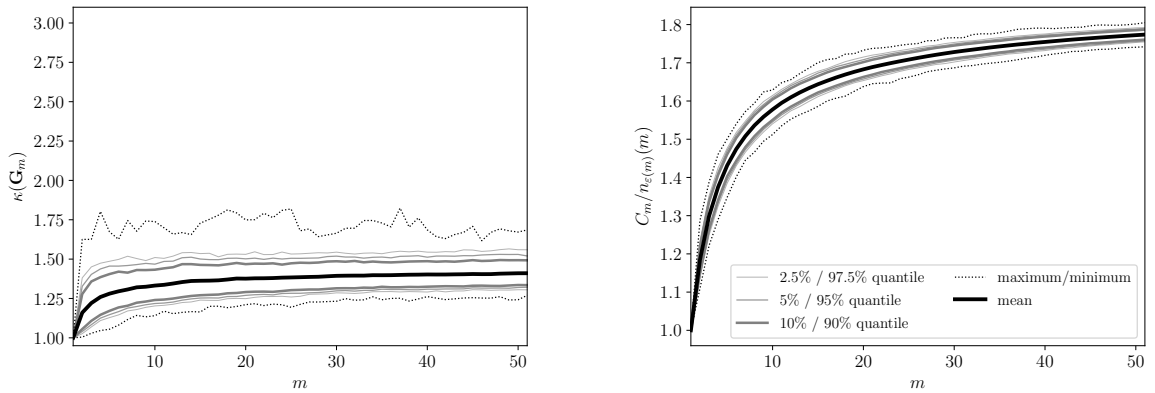


(a) Gramian condition numbers (b) Cost comparison

Figure 2. *Results for Algorithm 1 with* $n(m) = n_\varepsilon(m)$ *as in* (3.6), $\varepsilon = 10^{-2}$, *applied to Hermite polynomials of degrees* $0, \dots, m-1$.

each case, the quantiles of the corresponding distributions are estimated from 1000 test runs. Figure 2 shows the results for Algorithm 1 in case (i) with $n(m) = n_\varepsilon(m)$ as in (3.6), where we choose $\varepsilon = 10^{-2}$. We know from Theorem 2.1 that, for each m, one has $\kappa(\mathbf{G}_m) \leq 3$ with probability greater than $1 - \varepsilon$. In fact, this bound appears to be fairly pessimistic, since no sample with $\kappa(\mathbf{G}_m) > 3$ is encountered in the present test. In Figure 2(b), we show the ratio $C_m/n_\varepsilon(m)$. Recall that C_m is defined in (3.5) as the total number of samples used in the repeated application of Algorithm 1 to produce $S_1, \dots, S_m$, and thus $C_m/n_\varepsilon(m)$ provides a comparison to the costs of directly sampling S_m from μ_m. The results are in agreement with Theorems 3.2 and 3.4, which show in particular that $C_m < 2n_\varepsilon(m)$ with very high probability.

Using the same setup with $n(m) = n_{\varepsilon(m)}(m)$ as in (3.22), where $\varepsilon_0 = 10^{-2}$, leads to the expected improved uniformity in $\kappa(\mathbf{G}_m)$. The corresponding results are shown in Figure 3,



(a) Gramian condition numbers

(b) Cost comparison

Figure 3. *Results for Algorithm 1 with $n(m) = n_{\varepsilon(m)}(m)$ as in (3.22), $\varepsilon_0 = 10^{-2}$, applied to Hermite polynomials of degrees $0, \dots, m-1$.*

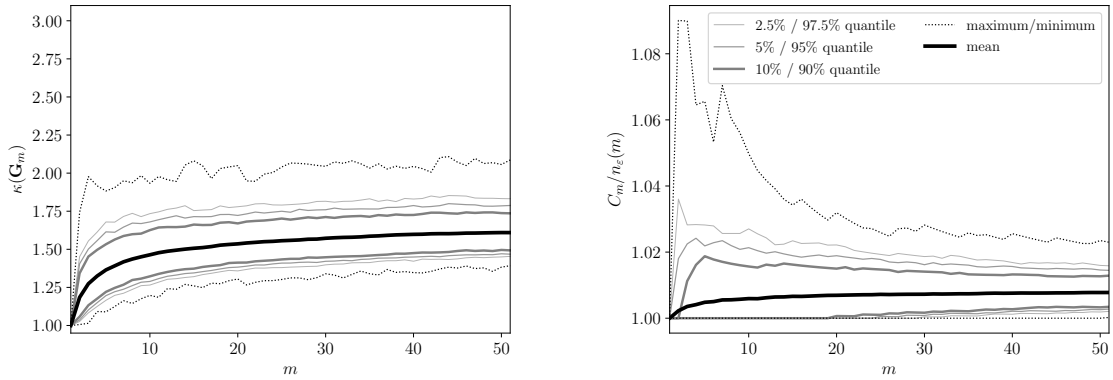
with sampling costs that are in agreement with (3.26) and Theorem 3.6. Since the effects of replacing $n_\varepsilon(m)$ by $n_{\varepsilon(m)}(m)$ are very similar in all further tests, we only show results for $n(m) = n_\varepsilon(m)$ with $\varepsilon = 10^{-2}$ in what follows.

While the simple scheme in Algorithm 1 already ensures near-optimal costs with high probability, there are some practical variants that can yield better quantitative performance. A first such variant is given in Algorithm 2. Instead of deciding for each previous sample separately whether it will be reused, here a queue of previous samples is kept, from which these are extracted in order until the previous sample set S_m is exhausted. Clearly, the costs of this scheme are bounded from above by those of Algorithm 1.

Algorithm 2. Sequential sampling with sample queue.

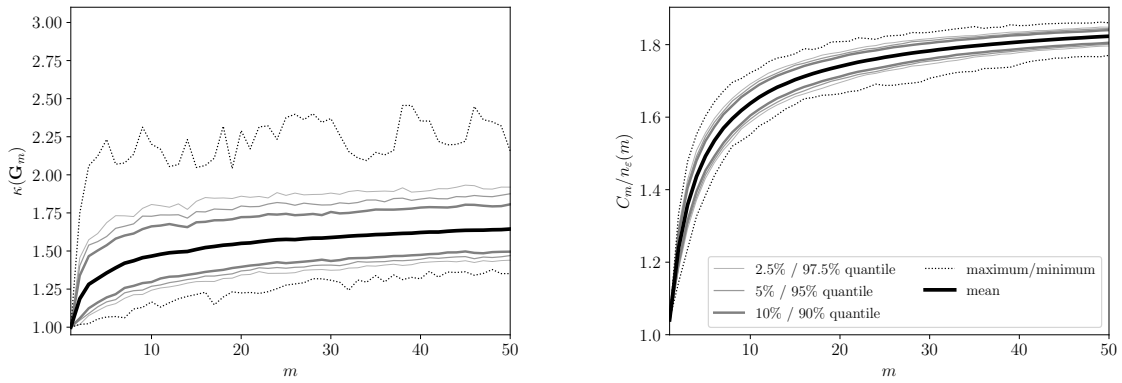
input: sample $S_m = \{x_m^1, \dots, x_m^{n(m)}\}$ from μ_m
output: sample $S_{m+1} = \{x_{m+1}^1, \dots, x_{m+1}^{n(m+1)}\}$ from μ_{m+1}

$j := 1$
for $i = 1, \dots, n(m+1)$ **do**
  draw a_i uniformly distributed in $\{1, \dots, m+1\}$
  if $a_i = m+1$ **then**
    draw x_{m+1}^i from σ_{m+1}
  else if $j \le n(m)$ **then**
    $x_{m+1}^i := x_m^j$
    $j \leftarrow j+1$
  else
    draw x_{m+1}^i from μ_m by (3.3)
  end if
end for



(a) Gramian condition numbers (b) Cost comparison

Figure 4. *Results for Algorithm 2 with $n(m) = n_\varepsilon(m)$ as in (3.6), $\varepsilon = 10^{-2}$, applied to Hermite polynomials of degrees $0, \dots, m-1$.*



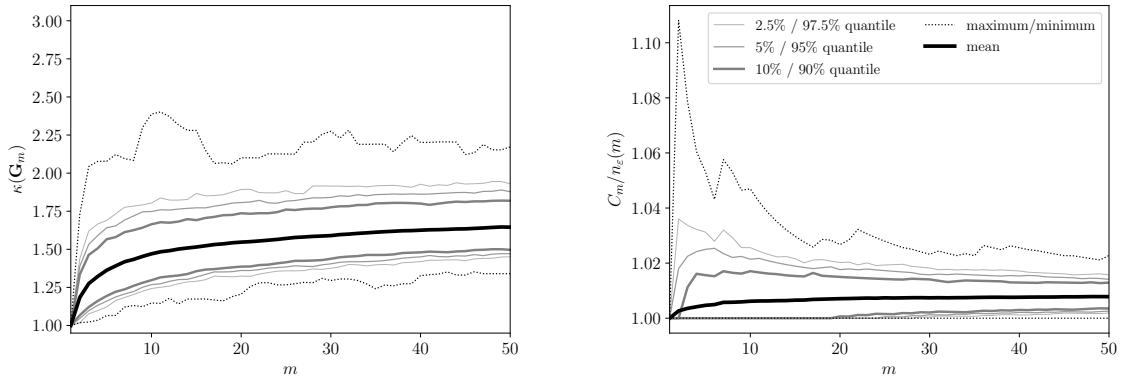
(a) Gramian condition numbers (b) Cost comparison

Figure 5. *Results for Algorithm 1 with $n(m) = n_\varepsilon(m)$ as in (3.6), $\varepsilon = 10^{-2}$, applied to a subset of Haar basis obtained by random tree refinement.*

As expected, in the results for Algorithm 2 applied in case (i), which are shown in Figure 4, we find an estimate of the distribution of $\kappa(\mathbf{G}_m)$ that is essentially identical to the one for Algorithm 1 in Figure 2(a). The costs, however, are substantially more favorable than the ones in Figure 2(b): using Algorithm 2, the successive sampling of $S_1, \dots, S_m$ uses only a small fraction of additional samples when compared to directly sampling only S_m.

Figures 5 and 6 show the analogous comparison of Algorithms 1 and 2 applied to case (ii), which leads to very similar results. This is not surprising, considering that the bounds in the general Theorem 2.1 on optimal least squares sampling, as well as those in section 3, are all independent of the chosen L^2-space and of the corresponding orthonormal basis. Our numerical results thus indicate that one can indeed expect a rather minor effect of this choice also in practice.

A further algorithmic variant consists in applying the inner loop of Algorithm 2 until



(a) Gramian condition numbers

(b) Cost comparison

Figure 6. *Results for Algorithm 2 with $n(m) = n_\varepsilon(m)$ as in (3.6), $\varepsilon = 10^{-2}$, applied to a subset of Haar basis obtained by random tree refinement.*

Algorithm 3. Sequential sampling with guaranteed condition number.

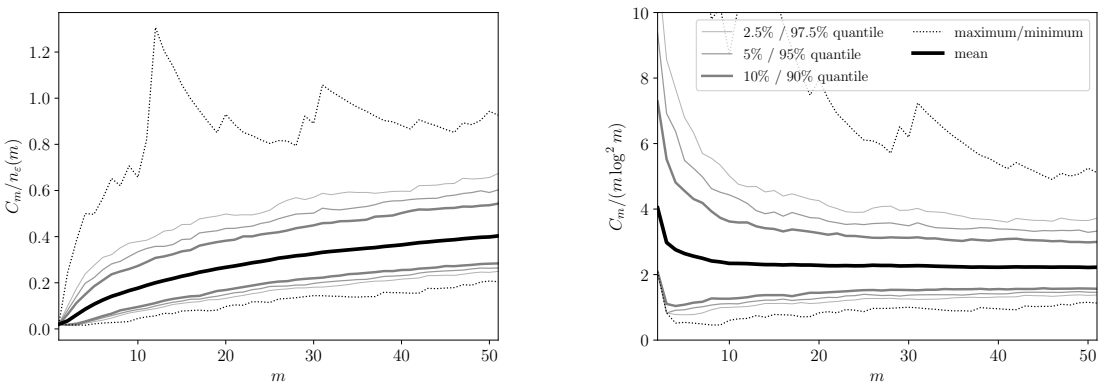
input: sample $S_m = \{x_m^1, \dots, x_m^{\hat{n}_m}\}$ from μ_m

output: sample $S_{m+1} = \{x_{m+1}^1, \dots, x_{m+1}^{\hat{n}_{m+1}}\}$ from μ_{m+1}

$j := 1$, $i := 1$, $\lambda := 1$
repeat
  draw a_i uniformly distributed in $\{1, \dots, m+1\}$
  if $a_i = m+1$ **then**
    draw x_{m+1}^i from σ_{m+1}
  else if $j \le \hat{n}_m$ **then**
    $x_{m+1}^i := x_m^j$
    $j \leftarrow j + 1$
  else
    draw x_{m+1}^i from μ_m by (3.3)
  end if
  if $i \ge m+1$ **then**
    Assemble $\tilde{\mathbf{G}}_{m+1}^i$ according to (2.4) using $\{x_{m+1}^1, \dots, x_{m+1}^i\}$
    $\lambda \leftarrow \|\tilde{\mathbf{G}}_{m+1}^i - \mathbf{I}\|$
  end if
  $i \leftarrow i + 1$
until $\lambda \le \frac{1}{2}$

ensuring that the stability criterion $\|\mathbf{G}_m - \mathbf{I}\| \le \frac{1}{2}$ is met, so that in particular $\kappa(\mathbf{G}_m) \le 3$ holds with certainty. This procedure is described in Algorithm 3. Note that here the size $\hat{n}_m$ of the sample S_m is not fixed a priori.

Since the stability criterion is ensured with certainty by this third algorithm, we only need to study the total sampling cost C_m. This is illustrated in Figure 7 in case (i) of Hermite



(a) Ratio between C_m and $n_\varepsilon(m)$ (b) Ratio between C_m and $m\log^2 m$.

Figure 7. *Results for Algorithm 3 applied to Hermite polynomials of degrees* $0, \dots, m-1$.

polynomials. For a direct comparison to Algorithms 1 and 2, we compare the costs to $n_\varepsilon(m)$ as before, although this value plays no role in Algorithm 3. Figure 7(a) shows that with high probability, one has $C_m/n_\varepsilon(m) < 1$ for the considered range of m and $\varepsilon = 10^{-2}$, although this ratio can be seen to increase approximately logarithmically. A closer inspection shows that C_m tends to behave like $m\log^2 m$, as illustrated in Figure 7(b). This hints that an extra logarithmic factor is needed for ensuring stability with certainty for all values of m.

Acknowledgment. The authors would like to thank Sören Wolfers for pointing out a correction.

REFERENCES

[1] D. Angluin and L.G. Valiant, *Fast probabilistic algorithms for Hamiltonian circuits and mappings*, J. Comput. System Sci., 18 (1979), pp. 155–193.
[2] A. Chkifa, A. Cohen, G. Migliorati, F. Nobile, and R. Tempone, *Discrete least squares polynomial approximation with random evaluations—application to parametric and stochastic PDEs*, ESAIM Math. Model. Numer. Anal., 49 (2015), pp. 815–837.
[3] A. Cohen, *Numerical Analysis of Wavelet Methods*, Stud. Math. Appl., Elsevier, Amsterdam, 2003.
[4] A. Cohen and R. DeVore, *Approximation of high-dimensional PDEs*, Acta Numer., 24 (2015), pp. 1–159.
[5] A. Cohen, R. DeVore, and C. Schwab, *Analytic regularity and polynomial approximation of parametric and stochastic PDEs*, Anal. Appl., 9 (2011), pp. 11–47.
[6] A. Cohen and G. Migliorati, *Optimal weighted least squares methods*, SMAI J. Comput. Math., 3 (2017), pp. 181–203.
[7] 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. Vols. 1 and 2, Springer, Cham, 2018, pp. 233–282.
[8] F. Cucker and S. Smale, *On the mathematical foundations of learning*, Bull. Amer. Math. Soc. (N.S.), 39 (2001), pp. 1–49.
[9] W. Dahmen, *Wavelet and multiscale methods for operator equations*, Acta Numer., 6 (1997), pp. 55–228.
[10] R. DeVore, *Nonlinear approximation*, Acta Numer., 7 (1998), pp. 51–150.
[11] A. Doostan and M. Hadigol, *Least squares polynomial chaos expansion: A review of sampling strategies*, Comput. Methods Appl. Mech. Engrg., 332 (2018), pp. 382–407.

[12] 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.
[13] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer, Berlin, 2002.
[14] T. Hagerup and C. Rüb, *A guided tour of Chernoff bounds*, Inform. Process. Lett., 33 (1990), pp. 305–308.
[15] A.L. Haji-Ali, F. Nobile, R. Tempone, and S. Wolfers, *Multilevel Weighted Least Squares Polynomial Approximation*, preprint, https://arxiv.org/abs/1707.00026, 2017.
[16] J. Hampton and A. Doostan, *Basis adaptive sample efficient polynomial chaos (BASE-PC)*, J. Comput. Phys., 371 (2018), pp. 20–49.
[17] 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.
[18] G. Mastroianni and V. Totik, *Weighted polynomial inequalities with doubling and a_∞ weights*, Constr. Approx., 16 (2000), pp. 37–71.
[19] G. Migliorati, *Adaptive Approximation by Optimal Weighted Least Squares Methods*, preprint, https://arxiv.org/abs/1807.00402, 2018.
[20] P. Nevai, T. Erdélyi, and A.P. Magnus, *Generalized Jacobi weights, Christoffel functions, and Jacobi polynomials*, SIAM J. Math. Anal., 25 (1994), pp. 602–614, https://doi.org/10.1137/S0036141092236863.
[21] J. Tropp, *User-friendly tail bounds for sums of random matrices*, Found. Comput. Math., 12 (2012), pp. 389–434.