

Boosted optimal weighted least-squares

Cécile Haberstich^{*†}Anthony Nouy[†]Guillaume Perrin^{*}

Abstract

This paper is concerned with the approximation of a function u in a given approximation space V_m of dimension m from evaluations of the function at n suitably chosen points. The aim is to construct an approximation of u in V_m which yields an error close to the best approximation error in V_m and using as few evaluations as possible. Classical least-squares regression, which defines a projection in V_m from n random points, usually requires a large n to guarantee a stable approximation and an error close to the best approximation error. This is a major drawback for applications where u is expensive to evaluate. One remedy is to use a weighted least squares projection using n samples drawn from a properly selected distribution. In this paper, we introduce a boosted weighted least-squares method which allows to ensure almost surely the stability of the weighted least squares projection with a sample size close to the interpolation regime $n = m$. It consists in sampling according to a measure associated with the optimization of a stability criterion over a collection of independent n -samples, and resampling according to this measure until a stability condition is satisfied. A greedy method is then proposed to remove points from the obtained sample. Quasi-optimality properties are obtained for the weighted least-squares projection, with or without the greedy procedure. The proposed method is validated on numerical examples and compared to state-of-the-art interpolation and weighted least squares methods.

Keywords— approximation, weighted least-squares, optimal sampling, error analysis, greedy algorithm, interpolation

1 Introduction

The continuous improvement of computational resources makes the role of the numerical simulation always more important for modelling complex systems. However most of these numerical simulations remain very costly from a computational point of view. Furthermore, for many problems such as optimization, estimation or uncertainty quantification, the model is a function of possibly numerous parameters (design variables, uncertain parameters...) and has to be evaluated for many instances of the parameters. One remedy is to build an approximation of this function of the parameters which is further used as a surrogate model, or as a companion model used as a low-fidelity model.

This paper is concerned with the approximation of a function u using evaluations of the function at suitably chosen points. We consider functions from $L^2_\mu(\mathcal{X})$, the space of square-integrable functions defined on a set \mathcal{X} equipped with a probability measure μ . Given an approximation space V_m of dimension m in $L^2_\mu(\mathcal{X})$, the aim is to construct an approximation of u in V_m which yields an error close to the best approximation error in V_m and using as

^{*}CEA, DAM, DIF, F-91297 Arpajon France

[†]Centrale Nantes, LMJL UMR CNRS 6629, France

few evaluations as possible. A classical approach is least-squares regression, which defines the approximation by solving

$$\min_{v \in V_m} \frac{1}{n} \sum_{i=1}^n (u(x^i) - v(x^i))^2,$$

where the x^i are i.i.d. samples drawn from the measure μ . However, to guarantee a stable approximation and an error close to the best approximation error, least-squares regression may require a sample size n much higher than m (see [3]). This issue can be overcome by weighted least-squares projection, which is obtained by solving

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

where the x^i are points not necessarily drawn from μ and the $w(x^i)$ are corresponding weights. A suitable choice of weights and points may allow to decrease the sample size to reach the same approximation error, see e.g. [4, 6]. In [2], the authors introduce an optimal sampling measure ρ with a density $w(x)^{-1}$ with respect to the reference measure μ which depends on the approximation space. Choosing i.i.d. samples x^i from this optimal measure, one obtains with high probability $1 - \eta$ a stable approximation and an error of the order of the best approximation error using a sample size n in $O(m \log(m\eta^{-1}))$. Nevertheless, the necessary condition for having stability requires a sample size n much higher than m , especially when a small probability η is desired.

Here we introduce a boosted least-squares method which enables us to ensure almost surely the stability of the weighted least squares projection with a sample size tending to the interpolation regime $n = m$. It consists in sampling according to a measure associated with the optimization of a stability criterion over a collection of independent n -samples, and resampling according to this measure until a stability condition is satisfied. A greedy method is then proposed to remove points from the obtained sample. Quasi-optimality properties in expectation are obtained for the weighted least-squares projection, with or without the greedy procedure.

The outline of the paper is as follows. In section 2, we introduce the theoretical framework, and present some useful results on weighted least-squares projections. We recall the optimal sampling measure from [2], and outline its limitations. In section 3, we present the boosted least-squares approach. In section 4, we present numerical examples.

2 Least-squares method

Let \mathcal{X} be a subset of \mathbb{R}^d equipped with a probability measure μ , with $d \geq 1$. We consider a function u from $L_\mu^2(\mathcal{X})$, the Hilbert space of square-integrable real-valued functions defined on \mathcal{X} . We let $\|\cdot\|_{L_\mu^2}$ be the natural norm in $L_\mu^2(\mathcal{X})$ defined by

$$\|v\|_{L_\mu^2}^2 = \int_{\mathcal{X}} v(x)^2 d\mu(x). \quad (1)$$

When there is no ambiguity, $L_\mu^2(\mathcal{X})$ will be simply denoted L_μ^2 , and the norm $\|v\|_{L_\mu^2}$ and associated inner product $(\cdot, \cdot)_{L_\mu^2}$ will be denoted $\|\cdot\|$ and (\cdot, \cdot) respectively. Let V_m be a m -dimensional subspace of L_μ^2 , with $m \geq 1$, and $\{\varphi_j\}_{j=1}^m$ be an orthonormal basis of V_m . The best approximation of u in V_m is given by its orthogonal projection defined by

$$P_{V_m} u := \arg \min_{v \in V_m} \|u - v\|. \quad (2)$$

2.1 Weighted least-squares projection

Letting $\mathbf{x}^n := \{x^i\}_{i=1}^n$ be a set of n points in \mathcal{X} , we consider the weighted least-squares projection defined by

$$Q_{V_m}^{\mathbf{x}^n} u := \arg \min_{v \in V_m} \|u - v\|_{\mathbf{x}^n}, \quad (3)$$

where $\|\cdot\|_{\mathbf{x}^n}$ is a discrete semi-norm defined for v in L_μ^2 by

$$\|v\|_{\mathbf{x}^n}^2 := \frac{1}{n} \sum_{i=1}^n w(x^i) v(x^i)^2, \quad (4)$$

where w is a given non negative function defined on \mathcal{X} . We denote by

$$\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_m) : \mathcal{X} \rightarrow \mathbb{R}^m$$

the m -dimensional vector-valued function such that $\boldsymbol{\varphi}(x) = (\varphi_1(x), \dots, \varphi_m(x))^T$, and by $\mathbf{G}_{\mathbf{x}^n}$ the empirical Gram matrix defined by

$$\mathbf{G}_{\mathbf{x}^n} = \frac{1}{n} \sum_{i=1}^n w(x^i) \boldsymbol{\varphi}(x^i) \otimes \boldsymbol{\varphi}(x^i). \quad (5)$$

The stability of the weighted least-squares projection can be characterized by

$$Z_{\mathbf{x}^n} = \|\mathbf{G}_{\mathbf{x}^n} - \mathbf{I}\|_2,$$

which measures a distance between the empirical Gram matrix and the identity matrix \mathbf{I} , with $\|\cdot\|_2$ the matrix spectral norm. For any v in V_m , we have

$$(1 - Z_{\mathbf{x}^n})\|v\|^2 \leq \|v\|_{\mathbf{x}^n}^2 \leq (1 + Z_{\mathbf{x}^n})\|v\|^2. \quad (6)$$

We have the following properties that will be useful in subsequent analyses.

Lemma 2.1. *Let \mathbf{x}^n be a set of n points in \mathcal{X} such that $Z_{\mathbf{x}^n} = \|\mathbf{G}_{\mathbf{x}^n} - \mathbf{I}\|_2 \leq \delta$ for some $\delta \in [0, 1)$. Then*

$$(1 - \delta)\|v\|^2 \leq \|v\|_{\mathbf{x}^n}^2 \leq (1 + \delta)\|v\|^2 \quad (7)$$

and the weighted least-squares projection $Q_{V_m}^{\mathbf{x}^n} u$ associated with \mathbf{x}^n satisfies

$$\|u - Q_{V_m}^{\mathbf{x}^n} u\|^2 \leq \|u - P_{V_m} u\|^2 + (1 - \delta)^{-1} \|u - P_{V_m} u\|_{\mathbf{x}^n}^2. \quad (8)$$

Proof. The property (7) directly follows from (6) and $Z_{\mathbf{x}^n} \leq \delta$. Using the property of the orthogonal projection $P_{V_m} u$ and (7), we have that

$$\begin{aligned} \|u - Q_{V_m}^{\mathbf{x}^n} u\|^2 &= \|u - P_{V_m} u\|^2 + \|P_{V_m} u - Q_{V_m}^{\mathbf{x}^n} u\|^2 \\ &\leq \|u - P_{V_m} u\|^2 + (1 - \delta)^{-1} \|P_{V_m} u - Q_{V_m}^{\mathbf{x}^n} u\|_{\mathbf{x}^n}^2. \end{aligned}$$

Using the fact that $Q_{V_m}^{\mathbf{x}^n}$ is an orthogonal projection on V_m with respect to the semi-norm $\|\cdot\|_{\mathbf{x}^n}$, we have that for any v , $\|Q_{V_m}^{\mathbf{x}^n} v\|_{\mathbf{x}^n} \leq \|v\|_{\mathbf{x}^n}$. We deduce that

$$\|P_{V_m} u - Q_{V_m}^{\mathbf{x}^n} u\|_{\mathbf{x}^n} = \|Q_{V_m}^{\mathbf{x}^n} (P_{V_m} u - u)\|_{\mathbf{x}^n} \leq \|P_{V_m} u - u\|_{\mathbf{x}^n},$$

from which we deduce (8). \square

We now provide a result which bounds the L^2 error by a best approximation error with respect to a weighted supremum norm.

Theorem 2.2. Let \mathbf{x}^n be a set of n points in \mathcal{X} such that $Z_{\mathbf{x}^n} = \|\mathbf{G}_{\mathbf{x}^n} - \mathbf{I}\|_2 \leq \delta$ for some $\delta \in [0, 1)$. Then,

$$\|u - Q_{V_m}^{\mathbf{x}^n} u\| \leq (B + (1 - \delta)^{-1/2}) \inf_{v \in V_m} \|u - v\|_{\infty, w} \quad (9)$$

where $B^2 = \int_{\mathcal{X}} w(x)^{-1} d\mu(x)$ and $\|v\|_{\infty, w} = \sup_{x \in \mathcal{X}} w(x)^{1/2} |v(x)|$.

Proof. Using Lemma 2.1 we note that for any $v \in V_m$,

$$\|u - Q_{V_m}^{\mathbf{x}^n} u\| \leq \|u - v\| + (1 - \delta)^{-1/2} \|v - Q_{V_m}^{\mathbf{x}^n} u\|_{\mathbf{x}^n},$$

and $\|v - Q_{V_m}^{\mathbf{x}^n} u\|_{\mathbf{x}^n} = \|Q_{V_m}^{\mathbf{x}^n}(v - u)\|_{\mathbf{x}^n} \leq \|u - v\|_{\mathbf{x}^n}$. We then conclude by using the inequalities $\|u - v\|_{\mathbf{x}^n} \leq \|u - v\|_{\infty, w}$ and $\|u - v\| \leq \left(\int_{\mathcal{X}} w(x)^{-1} d\mu(x)\right)^{1/2} \sup_{x \in \mathcal{X}} w(x)^{1/2} |u(x) - v(x)|$. \square

In the case where w^{-1} is the density of a probability measure with respect to μ , (which will be the case in the rest of the paper), the constant B from Theorem 2.2 is equal to 1.

2.2 Random sampling

We consider the measure ρ on \mathcal{X} with density w^{-1} with respect to μ , i.e. $d\rho = w^{-1}d\mu$. If the x^1, \dots, x^n are i.i.d. random variables drawn from the measure ρ , or equivalently if $\mathbf{x}^n = (x^1, \dots, x^n)$ is drawn from the product measure $\rho^{\otimes n} := \boldsymbol{\rho}^n$ on \mathcal{X}^n , then for any function v in L_μ^2 (not only those in V_m), we have

$$\mathbb{E}(\|v\|_{\mathbf{x}^n}^2) = \|v\|^2. \quad (10)$$

The condition (10) restricted to all functions $v \in V_m$ implies that the empirical Gram matrix $\mathbf{G}_{\mathbf{x}^n}$ satisfies

$$\mathbb{E}(\mathbf{G}_{\mathbf{x}^n}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(w(x^i) \boldsymbol{\varphi}(x^i) \otimes \boldsymbol{\varphi}(x^i)) = \mathbf{I}. \quad (11)$$

The random variable $Z_{\mathbf{x}^n} = \|\mathbf{G}_{\mathbf{x}^n} - \mathbf{I}\|_2$ quantifies how much the random matrix $\mathbf{G}_{\mathbf{x}^n}$ deviates from its expectation. For any $\delta \in [0, 1)$, if

$$\mathbb{P}(Z_{\mathbf{x}^n} > \delta) \leq \eta, \quad (12)$$

then for all $v \in V_m$, Eq. (7) holds with probability higher than $1 - \eta$. We directly conclude from Theorem 2.2 that the weighted least-squares projection $Q_{V_m}^{\mathbf{x}^n}$ satisfies (9) with probability higher than $1 - \eta$ (and $B = 1$).

Now, we present results in expectation which relate the L^2 -error with the best approximation in L_μ^2 . We have the following result from [2] for a conditional weighted least-squares projection, here stated in a slightly different form.

Theorem 2.3 ([2]). Let \mathbf{x}^n be drawn from the measure $\boldsymbol{\rho}^n$ and let $Q_{V_m}^{\mathbf{x}^n} u$ be the associated weighted least-squares projection of u . For any $\delta \in [0, 1)$ and $\eta \in [0, 1]$ such that (12) holds,

$$\mathbb{E}(\|u - Q_{V_m}^{\mathbf{x}^n, C} u\|^2) \leq (1 + (1 - \delta)^{-1}) \|u - P_{V_m} u\|^2 + \eta \|u\|^2, \quad (13)$$

where $Q_{V_m}^{\mathbf{x}^n, C} u = Q_{V_m}^{\mathbf{x}^n} u$ if $Z_{\mathbf{x}^n} \leq \delta$ and 0 otherwise.

Proof. We have

$$\mathbb{E}(\|u - Q_{V_m}^{\mathbf{x}^n, C} u\|^2) = \mathbb{E}(\|u - Q_{V_m}^{\mathbf{x}^n} u\|^2 \mathbf{1}_{Z_{\mathbf{x}^n} \leq \delta}) + \|u\|^2 \mathbb{E}(\mathbf{1}_{Z_{\mathbf{x}^n} > \delta}),$$

with $\mathbb{E}(\mathbb{1}_{Z_{\mathbf{x}^n} > \delta}) = \mathbb{P}(Z_{\mathbf{x}^n} > \delta) \leq \eta$. Then using Lemma 2.1 and (10), we have

$$\begin{aligned} \mathbb{E}(\|u - Q_{V_m}^{\mathbf{x}^n} u\|^2 \mathbb{1}_{Z_{\mathbf{x}^n} \leq \delta}) &\leq \mathbb{E}(\|u - P_{V_m} u\|^2 + (1 - \delta)^{-1} \|u - P_{V_m} u\|_{\mathbf{x}^n}^2 \mathbb{1}_{Z_{\mathbf{x}^n} \leq \delta}) \\ &\leq \|u - P_{V_m} u\|^2 + (1 - \delta)^{-1} \mathbb{E}(\|u - P_{V_m} u\|_{\mathbf{x}^n}^2) \\ &= (1 + (1 - \delta)^{-1}) \|u - P_{V_m} u\|^2, \end{aligned}$$

which concludes the proof. \square

Also, we have the following quasi-optimality property for the weighted least-squares projection associated with the distribution $\boldsymbol{\rho}^n$ conditioned to the event $\{Z_{\mathbf{x}^n} \leq \delta\}$.

Theorem 2.4. *Let \mathbf{x}^n be drawn from the measure $\boldsymbol{\rho}^n$ and let $Q_{V_m}^{\mathbf{x}^n} u$ be the associated weighted least-squares projection of u . For any $\delta \in [0, 1)$ and $\eta \in [0, 1)$ such that (12) holds,*

$$\mathbb{E}(\|u - Q_{V_m}^{\mathbf{x}^n} u\|^2 | Z_{\mathbf{x}^n} \leq \delta) \leq (1 + (1 - \delta)^{-1} (1 - \eta)^{-1}) \|u - P_{V_m} u\|^2. \quad (14)$$

Proof. From Lemma (2.1), we have that

$$\mathbb{E}(\|u - Q_{V_m}^{\mathbf{x}^n} u\|^2 | Z_{\mathbf{x}^n} \leq \delta) \leq \|u - P_{V_m} u\|^2 + (1 - \delta)^{-1} \mathbb{E}(\|u - P_{V_m} u\|_{\mathbf{x}^n}^2 | Z_{\mathbf{x}^n} \leq \delta),$$

and

$$\mathbb{E}(\|u - P_{V_m} u\|_{\mathbf{x}^n}^2 | Z_{\mathbf{x}^n} \leq \delta) \leq \mathbb{E}(\|u - P_{V_m} u\|_{\mathbf{x}^n}^2) \mathbb{P}(Z_{\mathbf{x}^n} \leq \delta)^{-1},$$

and we conclude by using $\mathbb{P}(Z_{\mathbf{x}^n} \leq \delta) \geq 1 - \eta$ and the property (10). \square

2.3 Optimal sampling measure

An inequality of the form (12) can be obtained by concentration inequalities. A suitable sampling distribution can then be obtained by an optimization of the obtained upper bound. An optimal choice for w based on matrix Chernoff inequality is derived in [2] and given by

$$w(x)^{-1} = \frac{1}{m} \sum_{j=1}^m \varphi_j(x)^2 = \frac{1}{m} \|\boldsymbol{\varphi}(x)\|_2^2. \quad (15)$$

Using this distribution, we obtain the following result, for which we provide a sketch of proof following [2]. The result is here provided in a slightly more general form than in [2].

Theorem 2.5. *Let $\eta \in [0, 1)$ and $\delta \in [0, 1)$. Assume \mathbf{x}^n is drawn from the product measure $\boldsymbol{\rho}^n = \rho^{\otimes n}$, with ρ having the density (15) with respect to μ . If the sample size n is such that¹*

$$n \geq n(\delta, \eta, m) := d_\delta^{-1} m \log(2m\eta^{-1}), \quad (16)$$

with $d_\delta := -\delta + (1 + \delta) \log(1 + \delta)$, then $Z_{\mathbf{x}^n} = \|\mathbf{G}_{\mathbf{x}^n} - \mathbf{I}\|_2$ satisfies (12).

Proof. We have $\mathbf{G}_{\mathbf{x}^n} = \frac{1}{n} \sum_{i=1}^n \mathbf{A}_i$ where the $\mathbf{A}_i = w(x^i) \boldsymbol{\varphi}(x^i) \otimes \boldsymbol{\varphi}(x^i)$ are random matrices such that $\mathbb{E}(\mathbf{A}_i) = \mathbf{I}$ and $\|\mathbf{A}_i\|_2 = w(x^i) \|\boldsymbol{\varphi}(x^i)\|_2^2 = m$. The matrix Chernoff inequality from [7, Theorem 5.1] gives that the minimal and maximal eigenvalues of $\mathbf{G}_{\mathbf{x}^n} - \mathbf{I}$ satisfy

$$\mathbb{P}(\lambda_{\min}(\mathbf{G}_{\mathbf{x}^n} - \mathbf{I}) < -\delta) \vee \mathbb{P}(\lambda_{\max}(\mathbf{G}_{\mathbf{x}^n} - \mathbf{I}) > \delta) \leq m \exp(-nd_\delta/m).$$

Under the condition (16), we have that $m \exp(-nd_\delta/m) \leq \eta/2$ and using a union bound, we deduce (12). \square

¹Note that the constant in the condition (16) differs from the one given in the reference [2] for $\delta = 1/2$, which was incorrect.

Remark 2.6. Note that $d_\delta \leq \delta^2$. Then a sufficient condition for satisfying the condition (16) is $n \geq \delta^{-2} m \log(2m\eta^{-1})$.

Remark 2.7. The quantile function of $Z_{\mathbf{x}^n}$ is defined for $t \in [0, 1]$ by $F_{Z_{\mathbf{x}^n}}^-(t) = \inf\{\delta : F_{Z_{\mathbf{x}^n}}(\delta) \geq t\}$. For given n and η , $F_{Z_{\mathbf{x}^n}}^-(1 - \eta)$ is the minimal δ such that (12) is satisfied. Denoting by $\delta_c(\eta, n) = \min\{\delta : n \geq n(\delta, \eta, m)\}$, we clearly have $F_{Z_{\mathbf{x}^n}}^-(1 - \eta) \leq \delta_c(\eta, n)$. The closer δ_c is from $F_{Z_{\mathbf{x}^n}}^-(1 - \eta)$, the sharper the condition on the sample size n is for satisfying (12).

Theorem 2.5 states that using the optimal sampling density (15), a stable projection of u is obtained with a sample size in $O(m \log(m\eta^{-1}))$ with high probability. Note that a small probability η , and therefore a large sample size n , may be required for controlling the term $\eta \|u\|^2$ in the error bound (13) for the conditional projection, or for obtaining a quasi-optimality property (14) in conditional expectation with a quasi-optimality constant close to $1 + (1 - \delta)^{-1}$. This will be improved in the next section by proposing a new distribution (obtained by resampling, conditioning and subsampling) allowing to obtain stability with very high probability and a moderate sample size.

3 Boosted optimal weighted least-squares method

We here propose an improved weighted least-squares method by proposing distributions over \mathcal{X}^n having better properties than $\boldsymbol{\rho}^n = \rho^{\otimes n}$. The function w defining the weighted least-squares projections will always be taken such that w^{-1} is the density of the optimal sampling measure ρ with respect to the reference measure μ .

3.1 Resampling and conditioning

The first improvement consists in drawing M independent samples $\{\mathbf{x}^{n,i}\}_{i=1}^M$, with $\mathbf{x}^{n,i} = (x^{1,i}, \dots, x^{n,i})$, from the distribution $\boldsymbol{\rho}^n$, and then in selecting a sample $\mathbf{x}^{n,\star}$ which satisfies

$$\|\mathbf{G}_{\mathbf{x}^{n,\star}} - \mathbf{I}\|_2 = \min_{1 \leq i \leq M} \|\mathbf{G}_{\mathbf{x}^{n,i}} - \mathbf{I}\|_2, \quad (17)$$

where $\mathbf{G}_{\mathbf{x}}$ denote the empirical Gram matrix associated with a sample \mathbf{x} in \mathcal{X}^n . If several samples $\mathbf{x}^{n,i}$ are solutions of the minimization problem, $\mathbf{x}^{n,\star}$ is selected at random among the minimizers. We denote by $\boldsymbol{\rho}^{n,\star}$ the probability measure of $\mathbf{x}^{n,\star}$. The probability that the stability condition $Z_{\mathbf{x}^{n,\star}} = \|\mathbf{G}_{\mathbf{x}^{n,\star}} - \mathbf{I}\|_2 \leq \delta$ is verified can be made arbitrarily high, playing on M , as it is shown in the following lemma (whose proof is trivial).

Lemma 3.1. For any $\delta \in [0, 1)$ and $\eta \in (0, 1)$, if n satisfies (16), then

$$\mathbb{P}(Z_{\mathbf{x}^{n,\star}} \leq \delta) \geq 1 - \eta^M. \quad (18)$$

Therefore, we can choose a probability η arbitrary close to 1, so that the condition (16) does not require a large sample size n , and still obtain the stability condition with a probability at least $1 - \eta^M$ which can be made arbitrarily close to 1 by choosing a sufficiently large M . Even if $\boldsymbol{\rho}^n$ has a product structure, for $M > 1$, the distribution $\boldsymbol{\rho}^{n,\star}$ does not have a product structure, i.e. the components of $\mathbf{x}^{n,\star} = (x^{1,\star}, \dots, x^{n,\star})$ are not independent, and does not satisfy the assumptions of Theorems 2.3 and 2.4. In particular $\mathbb{E}(\mathbf{G}_{\mathbf{x}^{n,\star}})$ may not be equal to \mathbf{I} and in general, $\mathbb{E}(\|v\|_{\mathbf{x}^{n,\star}}^2) \neq \|v\|^2$ for an arbitrary function v when $M > 1$. Therefore, a new analysis of the resulting weighted least-squares projection is required.

Remark 3.2. Note that since the function $\mathbf{x} \mapsto \|\mathbf{G}_{\mathbf{x}} - \mathbf{I}\|_2$ defined on \mathcal{X}^d is invariant through permutations of the components of \mathbf{x} , we have that the components of $\mathbf{x}^{n,*}$ have the same marginal distribution.

In order to ensure that the stability property is verified almost surely we consider a sample $\tilde{\mathbf{x}}^n$ from the distribution $\tilde{\rho}^n$ of $\mathbf{x}^{n,*}$ knowing the event

$$A_\delta = \{\|\mathbf{G}_{\mathbf{x}^{n,*}} - \mathbf{I}\|_2 \leq \delta\}, \quad (19)$$

which is such that for any function f , $\mathbb{E}(f(\tilde{\mathbf{x}}^n)) = \mathbb{E}(f(\mathbf{x}^{n,*})|A_\delta)$. A sample $\tilde{\mathbf{x}}^n$ from the distribution $\tilde{\rho}^n$ is obtained by a simple rejection method, which consists in drawing samples $\mathbf{x}^{n,*}$ from the distribution $\rho^{n,*}$ until A_δ is satisfied.

Remark 3.3. Let J be the number of trials necessary to get a sample $\mathbf{x}^{n,*}$ verifying the stability condition A_δ . This random variable J follows a geometric distribution with a probability of success $\mathbb{P}(A_\delta)$. Therefore J is almost surely finite and

$$\mathbb{P}(J \geq k) = (1 - \mathbb{P}(A_\delta))^k, \quad (20)$$

i.e. the probability to have J greater than k decreases exponentially with k .

Now we provide a result on the distribution of $\tilde{\mathbf{x}}^n$ which will be later used for the analysis of the corresponding least-squares projection.

Lemma 3.4. Let $\tilde{\mathbf{x}}^n$ be a sample following the distribution $\tilde{\rho}^n$, which is the distribution $\rho^{n,*}$ knowing the event A_δ defined by (19). Assume that n satisfies the condition (16) for some $\eta \in (0, 1)$ and $\delta \in (0, 1)$. Then for any function v in L_μ^2 and any $0 < \varepsilon \leq 1$,

$$\mathbb{E}(\|v\|_{\tilde{\mathbf{x}}^n}^2) \leq \frac{C(\varepsilon, M)}{1 - \eta^M} \mathbb{E}(\|v\|_{\mathbf{x}^n}^{2/\varepsilon})^\varepsilon, \quad (21)$$

with

$$C(\varepsilon, M) = M \frac{(1 - \varepsilon)^{1-\varepsilon}}{(M - \varepsilon)^{1-\varepsilon}} \leq M.$$

In particular, for $\varepsilon = 1$,

$$\mathbb{E}(\|v\|_{\tilde{\mathbf{x}}^n}^2) \leq \frac{M}{1 - \eta^M} \|v\|^2. \quad (22)$$

Also, if $\|v\|_{\infty, w} = \sup_{x \in \mathcal{X}} w(x)^{1/2} |v(x)| < \infty$,

$$\mathbb{E}(\|v\|_{\tilde{\mathbf{x}}^n}^2) \leq \frac{C(\varepsilon, M)}{1 - \eta^M} \|v\|_{\infty, w}^{2-2\varepsilon} \|v\|^2. \quad (23)$$

Proof. See appendix. □

Corollary 3.5. Let $\tilde{\mathbf{x}}^n$ be a sample following the distribution $\tilde{\rho}^n$ and assume that n satisfies the condition (16) for some $\eta \in (0, 1)$ and $\delta \in (0, 1)$. For any $v \in L_\mu^2$, the weighted least-squares projection $Q_{V_m}^{\tilde{\mathbf{x}}^n} v$ associated with the sample $\tilde{\mathbf{x}}^n$ satisfies

$$\mathbb{E}(\|Q_{V_m}^{\tilde{\mathbf{x}}^n} v\|^2) \leq \frac{(1 - \delta)^{-1} M}{1 - \eta^M} \|v\|^2. \quad (24)$$

Proof. Since $Q_{V_m}^{\tilde{\mathbf{x}}^n} v \in V_m$, we have that

$$\|Q_{V_m}^{\tilde{\mathbf{x}}^n} v\|^2 \leq (1 - \delta)^{-1} \|Q_{V_m}^{\tilde{\mathbf{x}}^n} v\|_{\tilde{\mathbf{x}}^n}^2 \leq (1 - \delta)^{-1} \|v\|_{\tilde{\mathbf{x}}^n}^2, \quad (25)$$

where we have used the fact that $Q_{V_m}^{\tilde{\mathbf{x}}^n}$ is an orthogonal projection with respect to the semi-norm $\|\cdot\|_{\tilde{\mathbf{x}}^n}$. Taking the expectation and using eq. (22), we obtain

$$\mathbb{E}(\|Q_{V_m}^{\tilde{\mathbf{x}}^n} v\|^2) \leq \frac{(1 - \delta)^{-1} M}{1 - \eta^M} \|v\|^2. \quad (26)$$

□

Theorem 3.6. *Let $\tilde{\mathbf{x}}^n$ be a sample following the distribution $\tilde{\rho}^n$ and assume that n satisfies the condition (16) for some $\eta \in (0, 1)$ and $\delta \in (0, 1)$. The weighted least-squares projection $Q_{V_m}^{\tilde{\mathbf{x}}^n} u$ associated with the sample $\tilde{\mathbf{x}}^n$ satisfies the quasi-optimality property*

$$\mathbb{E}(\|u - Q_{V_m}^{\tilde{\mathbf{x}}^n} u\|^2) \leq (1 + \frac{(1 - \delta)^{-1} M}{1 - \eta^M}) \|u - P_{V_m} u\|^2. \quad (27)$$

Also, assuming $\|u\|_{\infty, w} \leq L$, we have

$$\mathbb{E}(\|u - Q_{V_m}^{\tilde{\mathbf{x}}^n} u\|^2) \leq (1 + \frac{(1 - \delta)^{-1}}{1 - \eta^M} D(M, L, m)) \|u - P_{V_m} u\|^2 \quad (28)$$

where

$$D(M, L, m) = M \inf_{0 < \varepsilon \leq 1} C(\varepsilon, M) (L(1 + c_m))^{2-2\varepsilon},$$

with $C(\varepsilon, M)$ defined in Lemma 3.4 and c_m the supremum of $\|P_{V_m} v\|$ over functions v such that $\|v\|_{\infty, w} \leq 1$.

Proof. From Lemma 2.1, we have that

$$\|u - Q_{V_m}^{\tilde{\mathbf{x}}^n} u\|^2 \leq \|u - P_{V_m} u\|^2 + (1 - \delta)^{-1} \|u - P_{V_m} u\|_{\tilde{\mathbf{x}}^n}^2$$

holds almost surely, and from Lemma 3.4, we have that

$$\mathbb{E}(\|u - P_{V_m} u\|_{\tilde{\mathbf{x}}^n}^2) \leq \frac{C(\varepsilon, M)}{1 - \eta^M} \mathbb{E}(\|u - P_{V_m} u\|_{\tilde{\mathbf{x}}^n}^{2/\varepsilon})^\varepsilon$$

for all $\varepsilon \in (0, 1]$. Combining the above inequalities and then taking the infimum over ε , we obtain

$$\mathbb{E}(\|u - Q_{V_m}^{\tilde{\mathbf{x}}^n} u\|^2) \leq \|u - P_{V_m} u\|^2 + \frac{(1 - \delta)^{-1}}{1 - \eta^M} \inf_{0 < \varepsilon \leq 1} C(\varepsilon, M) \mathbb{E} \left(\|u - P_{V_m} u\|_{\tilde{\mathbf{x}}^n}^{2/\varepsilon} \right)^\varepsilon. \quad (29)$$

The particular case $\varepsilon = 1$ yields (27). The second property (28) is simply deduced from (29) by using the property (23) of Lemma 3.4 and by noting that $\|u - P_{V_m} u\|_{\infty, w} \leq (1 + c_m) \|u\|_{\infty, w}$. □

The quasi-optimality property (28) may improve (29) if $D(M, L, m) \leq M$, which may happen under some conditions on M, L, m .

Remark 3.7. *The constant c_m in Theorem 3.6 is such that $c_m \leq m$. Indeed, $P_{V_m} v(x) = \sum_{i=1}^m a_i \varphi_i(x)$ with*

$$a_i = (v, \varphi_i) = \int v(x) \varphi_i(x) d\mu(x) = \int v(x) \varphi_i(x) w(x) d\rho(x),$$

so that

$$|a_i| \leq \|v\|_{\infty,w} \int |\varphi_i(x)| w(x)^{1/2} d\rho(x) \leq \|v\|_{\infty,w} \left(\int \varphi_i(x)^2 w(x) d\rho(x) \right)^{1/2} = \|v\|_{\infty,w},$$

where we have used Cauchy-Schwarz inequality.

Therefore,

$$\begin{aligned} \|P_{V_m} v\|_{\infty,w} &\leq \|v\|_{\infty,w} \sup_{x \in \mathcal{X}} w(x)^{1/2} \sum_{i=1}^m |\varphi_i(x)| \\ &\leq \|v\|_{\infty,w} \sup_{x \in \mathcal{X}} w(x)^{1/2} m^{1/2} \left(\sum_{i=1}^m \varphi_i(x)^2 \right)^{1/2} = m \|v\|_{\infty,w}. \end{aligned}$$

3.2 Subsampling

Although the resampling enables us to choose δ and η such that n is smaller than with the initial strategy from [2], the value of n may still be high compared to an interpolation method. Therefore, to further decrease the sample size, for each generated sample $\tilde{\mathbf{x}}^n$, we propose to select a subsample which still verifies the stability condition.

We start with a sample $\tilde{\mathbf{x}}^n = (\tilde{x}^1, \dots, \tilde{x}^n)$ satisfying $\|\mathbf{G}_{\tilde{\mathbf{x}}^n} - \mathbf{I}\|_2 \leq \delta$ and then select a subsample $\tilde{\mathbf{x}}_K^n = (\tilde{x}^k)_{k \in K}$ with $K \subset \{1, \dots, n\}$ such that the empirical Gram matrix $\mathbf{G}_{\tilde{\mathbf{x}}_K^n} = \frac{1}{\#K} \sum_{k \in K} w(\tilde{x}^k) \boldsymbol{\varphi}(\tilde{x}^k) \otimes \boldsymbol{\varphi}(\tilde{x}^k)$ still satisfies

$$\|\mathbf{G}_{\tilde{\mathbf{x}}_K^n} - \mathbf{I}\|_2 \leq \delta.$$

In practice, the set K is constructed by a greedy procedure. We start with $K = \{1, \dots, n\}$. Then at each step of the greedy procedure, we select k^* in K such that

$$\|\mathbf{G}_{\tilde{\mathbf{x}}_{K \setminus \{k^*\}}^n} - \mathbf{I}\|_2 = \min_{k \in K} \|\mathbf{G}_{\tilde{\mathbf{x}}_{K \setminus \{k\}}^n} - \mathbf{I}\|_2. \quad (30)$$

If $\|\mathbf{G}_{\tilde{\mathbf{x}}_{K \setminus \{k^*\}}^n} - \mathbf{I}\|_2 \leq \delta$ and $\#K > n_{\min}$ then k^* is removed from K . Otherwise, the algorithm is stopped. We denote by $\tilde{\boldsymbol{\rho}}_K^n$ the distribution of the sample $\tilde{\mathbf{x}}_K^n$ produced by this greedy algorithm.

Theorem 3.8. Assume n satisfies the condition (16) for some $\eta \in (0, 1)$ and $\delta \in (0, 1)$, and let $\tilde{\mathbf{x}}_K^n$ be a sample produced by the greedy algorithm with $\#K \geq n_{\min}$. The weighted least-squares projection $Q_{V_m}^{\tilde{\mathbf{x}}_K^n} u$ associated with the sample $\tilde{\mathbf{x}}_K^n$ satisfies the quasi-optimality property

$$\mathbb{E}(\|u - Q_{V_m}^{\tilde{\mathbf{x}}_K^n} u\|^2) \leq \left(1 + \frac{n}{n_{\min}} \frac{(1 - \delta)^{-1}}{1 - \eta^M} M\right) \|u - P_{V_m} u\|^2. \quad (31)$$

Also, assuming $\|u\|_{\infty,w} \leq L$, we have

$$\mathbb{E}(\|u - Q_{V_m}^{\tilde{\mathbf{x}}_K^n} u\|^2) \leq \left(1 + \frac{n}{n_{\min}} \frac{(1 - \delta)^{-1}}{1 - \eta^M} D(M, L, m)\right) \|u - P_{V_m} u\|^2 \quad (32)$$

where $D(M, L, m)$ is defined in Theorem 3.8.

Proof. Since $Z_{\tilde{\mathbf{x}}_K^n} \leq \delta$, from Lemma 2.1, we have that for any $v \in V_m$, the least squares projection associated with $\tilde{\mathbf{x}}_K^n$ satisfies

$$\begin{aligned} \|u - Q_{V_m}^{\tilde{\mathbf{x}}_K^n} u\|^2 &\leq \|u - P_{V_m} u\|^2 + (1 - \delta)^{-1} \|u - P_{V_m} u\|_{\tilde{\mathbf{x}}_K^n}^2 \\ &\leq \|u - P_{V_m} u\|^2 + (1 - \delta)^{-1} \frac{n}{\#K} \|u - P_{V_m} u\|_{\tilde{\mathbf{x}}^n}^2, \end{aligned} \quad (33)$$

where the second inequality simply results from

$$\|v\|_{\tilde{\mathbf{x}}_K^n}^2 = \frac{1}{\#K} \sum_{k \in K} w(\tilde{x}^k) v(\tilde{x}^k)^2 \leq \frac{1}{\#K} \sum_{k=1}^n w(\tilde{x}^k) v(\tilde{x}^k)^2 = \frac{n}{\#K} \|v\|_{\tilde{\mathbf{x}}^n}.$$

Therefore, since $\#K \geq n_{\min}$, we obtain from Lemma 3.4 that

$$\mathbb{E}(\|u - Q_{V_m}^{\tilde{\mathbf{x}}_K^n} u\|^2) \leq \|u - P_{V_m} u\|^2 + \frac{n}{n_{\min}} \frac{(1 - \delta)^{-1}}{1 - \eta^M} \inf_{0 < \varepsilon \leq 1} C(\varepsilon, M) \mathbb{E} \left(\|u - P_{V_m} u\|_{\tilde{\mathbf{x}}^n}^{\frac{2}{\varepsilon}} \right)^\varepsilon.$$

The particular case $\varepsilon = 1$ yields the first property. For the second property, the proof follows the one of the property (28) in Theorem 3.6. \square

If we set $n_{\min} = m$, it may happen that the algorithm runs until $\#K = m$, the interpolation regime. Choosing $n \geq n(\delta, \eta, m)$ then yields a quasi-optimality constant depending on $\log(m)$. It has to be compared with the optimal behaviour of the Lebesgue constant for polynomial interpolation in one dimension. If we choose $n_{\min} = n/\beta$ for some fixed $\beta \geq 1$ independent of m , then we have $\frac{n}{n_{\min}} \leq \beta$ and a quasi-optimality constant independent of m in (3.2), but the algorithm may stop before reaching the interpolation regime ($n = m$).

4 Numerical experiments

4.1 Notations and objectives

In this section, we focus on polynomial approximation spaces $V_m = \mathbb{P}_p$ with p the polynomial degree. We use an orthonormal polynomial basis of V_m (Hermite polynomials for a Gaussian measure or Legendre polynomials for a uniform measure). The aim is to compare the performance of the method we propose with the optimal weighted least-squares method and interpolation. More precisely, we will compare the 4 following different approximation methods:

- interpolation performed on a deterministic set of points (Gauss-Hermite points for a Gaussian measure and Gauss-Legendre points for a uniform measure), simply denoted **I**,
- empirical interpolation, computed with magic points (see [5]) chosen among a large set of points randomly sampled from the measure μ , abbreviated **EI**,
- optimal weighted least-squares projection (introduced in [2]), abbreviated **OWLS**,
- the boosted optimal weighted least-squares projection we propose, abbreviated **BLS**, **c-BLS** and **s-BLS** when we respectively use resampling, conditioning, and subsampling plus conditioning.

Remark 4.1. For a fixed approximation space V_m , it must be noticed that the methods **OWLS**, **BLS** and **I** do not depend on the choice of the orthonormal basis associated with V_m , as the quantity $Z_{\mathbf{x}^n}$ is independent of this choice. This is however not the case for the **EI** method [5].

In the next section, two kinds of comparisons are performed. First, we compare qualitatively the distributions of the random variable $Z_{\mathbf{x}^n}$ and the distributions of the n -points sample \mathbf{x}^n . These analyses depend only on the choice of the approximation space V_m , and does not involve a function to approximate. Secondly, we compare quantitatively the efficiency of the different methods to approximate functions. We consider analytical functions on \mathbb{R} or $[-1, 1]$ equipped with Gaussian or uniform measures.

4.2 Qualitative analysis of the boosted optimal weighted least-squares method

4.2.1 Analysis of the stability

The objective of this paragraph is to compare the stability of the boosted optimal weighted least-squares method, using subsampling from section 3.2 or not, respectively **s-BLS** and **c-BLS**, with two other state-of-the art methods, standard least-squares method, abbreviated **SLS** and **OWLS** method. As explained in section 2.1, the stability of the least-squares projection can be characterized by the random variable $Z_{\mathbf{x}^n} = \|\mathbf{G}_{\mathbf{x}^n} - \mathbf{I}\|_2$. The closer $Z_{\mathbf{x}^n}$ is to 1, the more stable the approximation is. In this paragraph, we compare the distribution of this random variable $Z_{\mathbf{x}^n}$ for the different sampling methods. For the **SLS** method, the sampling measure is the reference measure μ . In the **OWLS** method, the sampling measure ρ is the measure with density w^{-1} with respect to the reference measure μ , chosen as in (15).

We present results for approximation spaces $V_m = \mathbb{P}_5$ with μ a Gaussian or uniform measure. The Figures 1a and 1b show that using **OWLS** instead of **SLS** shifts to the left the distribution of the random variable $Z_{\mathbf{x}^n}$. Without surprise, we see that conditioning $Z_{\mathbf{x}^n}$ by the event $A_\delta = \{Z_{\mathbf{x}^n} \leq \delta\}$ yields a distribution whose support is included in $[0, \delta]$. As expected, we also notice that very similar results are obtained for the **OWLS** and **c-BLS** methods when choosing $M = 1$. In the same manner, increasing the number of resampling M also shifts the PDF of $Z_{\mathbf{x}^n}$ to the low values, and decreases its variability. When interested in maximizing the probability of A_δ , **c-BLS** method is therefore an interesting alternative to **SLS** and **OWLS**. At last, looking at Figures 2a and 2b, we observe that the greedy selection moves the PDF of $Z_{\mathbf{x}^n}$ to the high values. This was expected: to switch from **c-BLS** to **s-BLS**, the size of the sample is reduced, as points are adaptively removed. However, as it is conditioned by A_δ , it remains better than **SLS** and **OWLS** methods.

4.2.2 Distribution of the sample points

In this paragraph, we are interested in the distributions of the points sampled with the **c-BLS** and **s-BLS** methods. We consider $d = 1$.

First, $n = 10$ points are sampled according the **c-BLS** method for different values of M (from 1 to 50000). These points are then sorted in ascending order. After repeating this procedure $r = 1000$ times, the probability distributions of the sorted points are represented in figs. 3 and 4 (one color per point).

For μ the Gaussian or the uniform measure, when M is small, ($M = 1$ or $M = 10$), we notice a strong overlap between the support of the different distributions. This is no longer the case for the highest values of M ($M = 10000$ or $M = 50000$). Hence, the larger M , the further apart the points are from each other with high probability, and the more they concentrate around specific values. Secondly, $n = 6$ points are sampled according the **s-BLS** method. To this end, a greedy procedure is applied to remove points from an initial sample of 10 points until we get the required number of points. In that case, as we fix the size of the sample, there is a priori no guaranty that the value of $Z_{\mathbf{x}^n}$ remains smaller than δ . The obtained 6-points sample is once again sorted in ascending order, and we repeat the procedure $r = 1000$ times. As previously, the distributions of the sorted points are represented in figs. 5 and 6 for different values of M . Only moderate values of M are considered, as we empirically observed that choosing M higher than 100 had very little influence on the results.

Comparing the figures associated with the methods with or without greedy subsampling, we

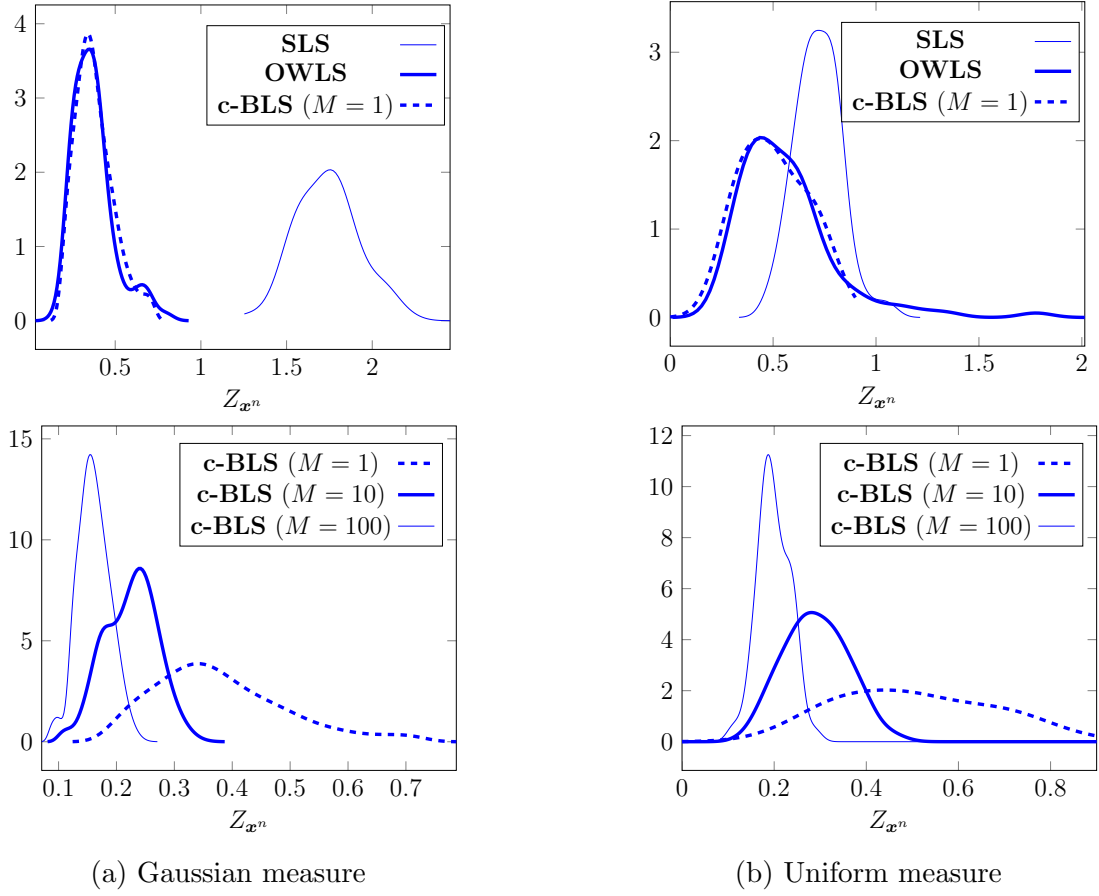


Figure 1: Probability density function of $Z_{x^n} = \|\mathbf{G}_{x^n} - \mathbf{I}\|_2$ for $V_m = \mathbb{P}_5$, with $\delta = 0.9$ and $n = 100$.

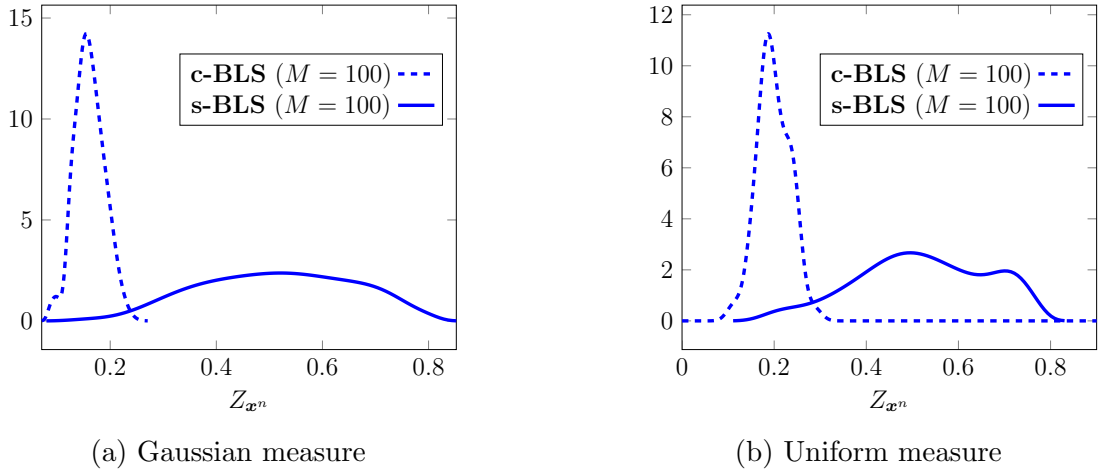


Figure 2: Probability density function of $Z_{x^n} = \|\mathbf{G}_{x^n} - \mathbf{I}\|_2$ for $V_m = \mathbb{P}_5$, with $\delta = 0.9$ and $n = 100$.

finally observe that **s-BLS** method provides results that are very close to **c-BLS** method with a very high value of M . This emphasizes the efficiency of the greedy selection to separate the support of the distributions of points.

In fig. 5a, the distributions of the sorted points associated to the **OWLS** method are represented in dashed lines. This shows that even if no resampling is carried out ($M = 1$), using the **s-BLS** method instead of **OWLS** improves the space filling properties of the obtained samples.

Remark 4.2. In fig. 3, fig. 4, fig. 5 and fig. 6 black dots have been added to indicate the positions

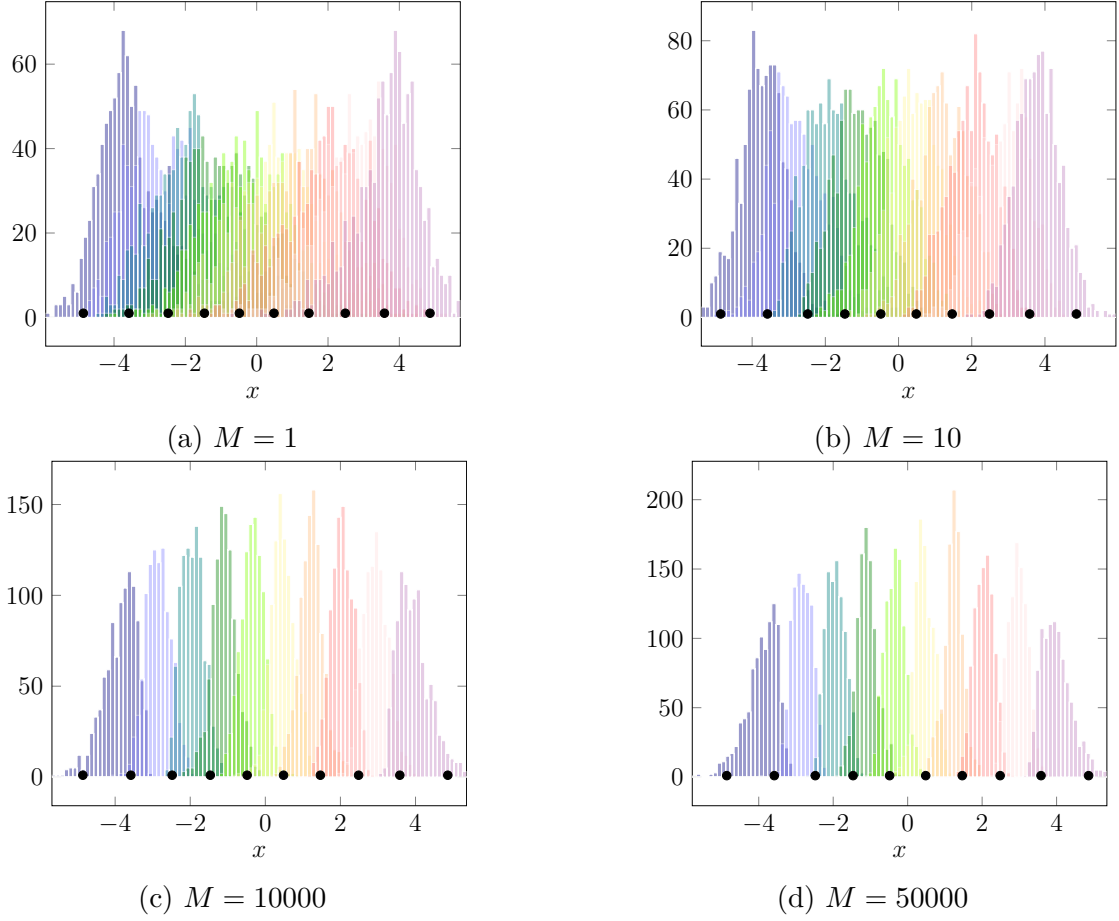


Figure 3: Distributions of the $x^{(i)}, i = 1, \dots, 10$, with \mathbf{x}^{10} sampled from the **c-BLS** method for $V_m = \mathbb{P}_5$ and μ the Gaussian measure.

of the first n Gauss-Hermite points in the Gaussian case, and the n first Gauss-Legendre points in the uniform case. Interestingly, we observe that, in the Gaussian case, the distribution of points spreads symmetrically around zero and in the uniform case, the distribution concentrates on the edges.

4.3 Quantitative analysis for polynomial approximations

In this paragraph, we want to compare the different methods introduced in section 4.1 in terms of approximation efficiency. The quality of the approximation u^* of a function $u \in L^2_\mu(\mathcal{X})$ is assessed by estimating the error of approximation $\varepsilon = \|u - u^*\|_{L^2_\mu(\mathcal{X})}$ with quadrature. Except for the deterministic interpolation method **I**, the points to compute the different approximations are drawn at random from a measure which depends on the approximation method that is considered. The different approximations are carried out 10 times (with different sets of points), and empirical confidence intervals of level 10% and 90% for the error of approximation are then computed.

For each example, two kind of comparisons are performed.

- In tables (a), we present results for the methods **OWLS** and **c-BLS**, for which the number of samples to ensure the stability of the approximation is given by Theorems 2.5 and 3.8.

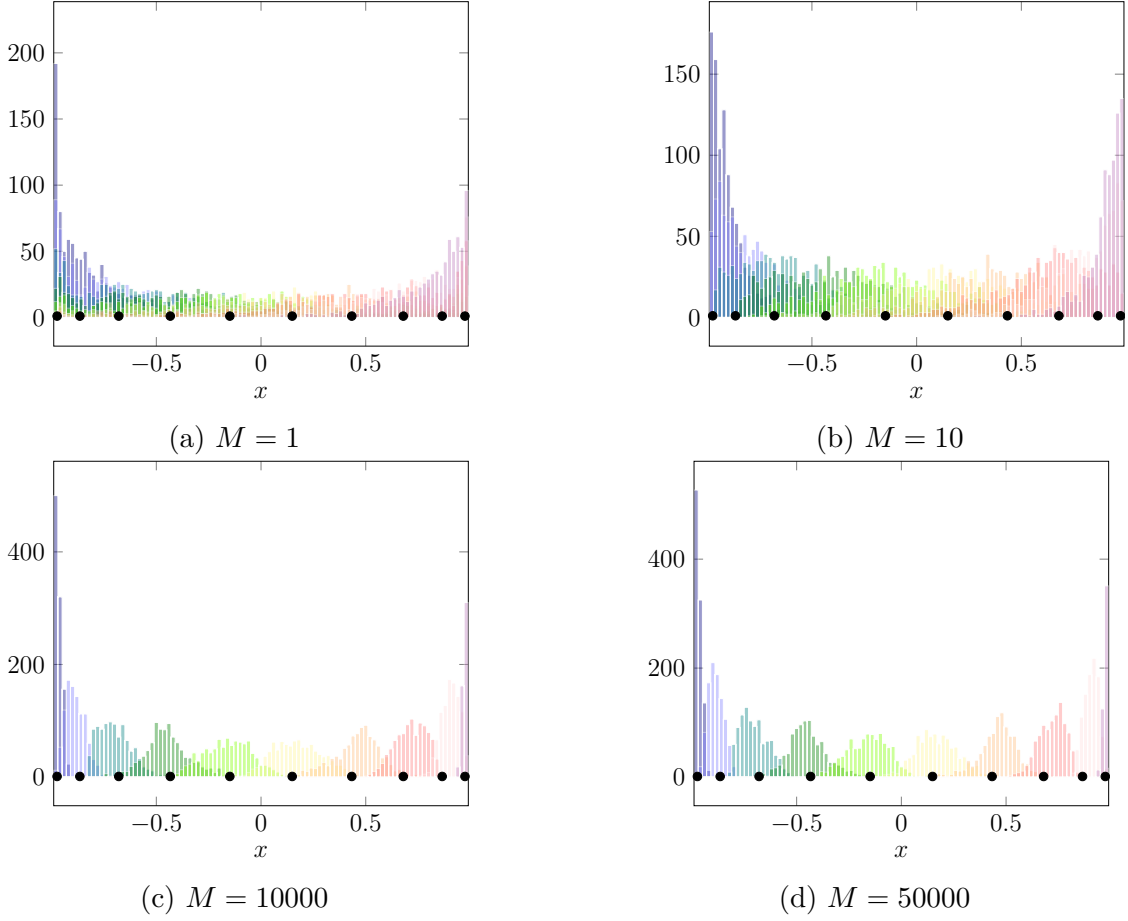


Figure 4: Distributions of the $x^{(i)}, i = 1, \dots, 10$, with \mathbf{x}^{10} sampled from the **c-BLS** method for $V_m = \mathbb{P}_5$ and μ the uniform measure.

- In tables (b), we compare the methods **I**, **EI**, **OWLS**, **BLS** and **s-BLS**. For the **I**, **EI**, **OWLS** and **BLS** methods, the number of samples n is taken equal to the dimension of the approximation space m . In this particular comparison, the **BLS** method only consists of a resampling strategy but without conditioning by the event $A_\delta = \{Z_{\mathbf{x}^n} \leq \delta\}$. For the **s-BLS** method, the initial number of samples n is taken as in eq. (16) and \mathbf{x}^n is conditioned by A_δ , such that the stability is guaranteed. Then the greedy selection of points is performed as long as the event A_δ is satisfied. In the examples presented in this paper, it leads to the interpolation regime $n = m$ for all trials.

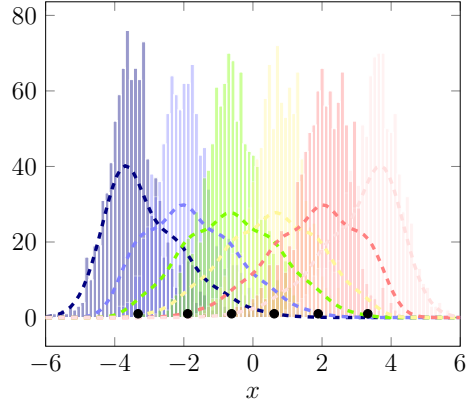
Remark 4.3. *In the interpolation regime $n = m$, the stability condition from eq. (12) can not be reached. Indeed, choosing M arbitrary big enables us to choose η close to 1, but still $\eta < 1$. It implies that in the number of samples $n(\delta, \eta, m)$ necessary to get the stability condition from Theorem 2.5 has to be greater than $d_\delta^{-1} m \log(2m) > m$. In the case of controlled cost, this explains why we choose to use the **BLS** method without conditioning.*

4.3.1 A first function

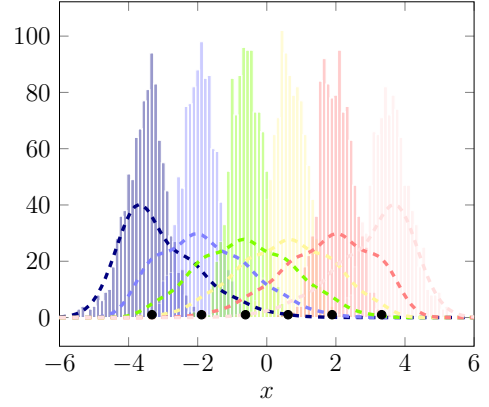
We consider $\mathcal{X} = \mathbb{R}$, equipped with the standard Gaussian measure μ and the function

$$u_1(x) = \exp\left(-\frac{1}{4}(x-1)^2\right). \quad (34)$$

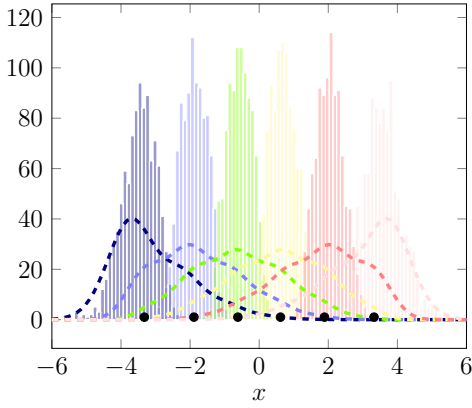
The approximation space is $V_m = \mathbb{P}_{m-1} = \text{span}\{\varphi_i : 1 \leq i \leq m\}$, with $\{\varphi_i\}_{i=1}^m$ the Hermite polynomials of degree less than $m-1$. This is referred to as example 1.



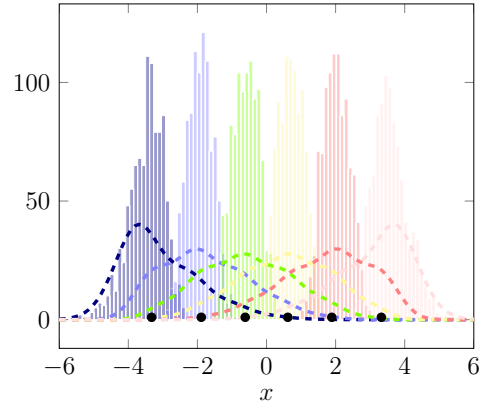
(a) $M = 1$



(b) $M = 10$



(c) $M = 50$



(d) $M = 100$

Figure 5: Distributions of the $x^{(i)}, i = 1, \dots, 6$, with \mathbf{x}^6 sampled from the **s-BLS** method (colored histograms) and **OWLS** method (dashed lines) for $V_m = \mathbb{P}_5$ with μ the gaussian measure.

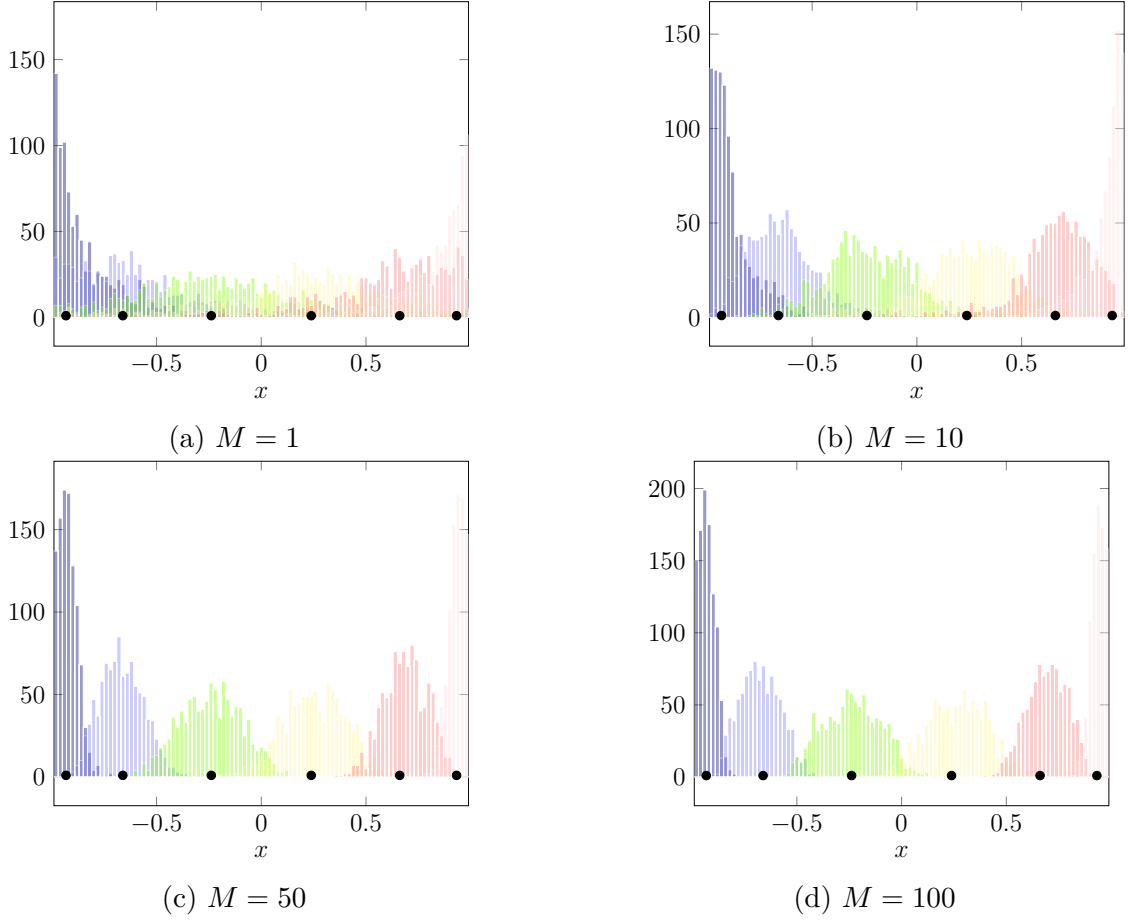


Figure 6: Distributions of the $x^{(i)}, i = 1, \dots, 6$, with \mathbf{x}^6 sampled from the **s-BLS** method for $V_m = \mathbb{P}_5$ and μ the uniform measure.

For this example, looking at table 1a, we first observe that the approximation error decreases in a similar way for the three methods **OWLS**, **c-BLS** ($M=1$), **c-BLS** ($M=100$), when the size of the approximation space increases. However, the results for the **c-BLS** ($M=100$) method are using less evaluations of the function. Indeed, by resampling, that is to say by increasing the value of M , the bound of the probability of getting a stable approximation is $1 - \eta^M$ instead of $1 - \eta$. Hence if η is chosen equal to 0.01 for $M = 1$, taking η equal to $0.01^{1/M}$ for higher values of M does not modify the bound of the probability of getting a stable approximation, but allows us to strongly reduce the number of samples needed to guarantee the same stability condition (see eq. (16) for the explicit relation between the minimum number of samples and η).

Looking at table 1b, we also observe that for all the methods, the error of approximation decreases when the size of the approximation space increases. Nevertheless, it is interesting to notice that the **EI** method faces numerical instabilities when the dimension of the approximation space is too high. In practice, we observe that for the **s-BLS** method, letting the greedy algorithm to reach the interpolation regime ($m = n$), the stability is verified for each of the 10 trials. However, imposing the number of samples n for the **s-BLS** method may provide a sample $\tilde{\mathbf{x}}^n$ which does not guarantee the stability. Focusing on the upper bound of the errors, we also see that in the interpolation regime, only the **s-BLS** method seems to be able to provide results that are compatible to the ones of the **I** method.

	OWLS		c-BLS ($M = 1$)		c-BLS ($M = 100$)	
	ε	n	ε	n	ε	n
$m = 6$	[-2.03; -2.03]	133	[-2.03; -2.02]	133	[-2.03; -2.01]	48
$m = 11$	[-3.03; -3.03]	265	[-3.03; -3.02]	265	[-3.03; -3.02]	108
$m = 16$	[-4.33; -4.33]	404	[-4.34; -4.3]	404	[-4.34; -4.31]	176
$m = 21$	[-5.67; -5.67]	548	[-5.67; -5.64]	548	[-5.65; -5.61]	249
$m = 26$	[-6.78; -6.78]	696	[-6.77; -6.74]	696	[-6.77; -6.73]	325
$m = 31$	[-8.27; -8.27]	847	[-8.26; -8.24]	847	[-8.26; -8.23]	405
$m = 36$	[-9.12; -9.12]	1001	[-9.11; -9.09]	1001	[-9.11; -9.07]	487
$m = 41$	[-10.48; -10.48]	1156	[-10.48; -10.43]	1156	[-10.48; -10.43]	571

(a) Guaranteed stability: the value of n is chosen to ensure a stable approximation with probability greater than 0.99.

	I	EI	OWLS	BLS ($M = 100$)	s-BLS
	ε	ε	ε	ε	ε
$m = 6$	-1.89	[-1.68; -1.32]	[-1.74; -1.1]	[-1.92; -1.7]	[-1.95; -1.73]
$m = 11$	-3.01	[-2.76; -2.44]	[-2.56; -0.71]	[-2.66; -2.29]	[-2.93; -2.7]
$m = 16$	-4.27	[-2.75; -2.4]	[-3.62; -1.48]	[-3.99; -3.55]	[-4.18; -3.98]
$m = 21$	-5.55	[-2.96; -2.66]	[-4.66; -2.73]	[-5.1; -4.61]	[-5.34; -5.09]
$m = 26$	-6.73	[-2.76; -2.44]	[-5.22; -3.68]	[-5.94; -5.22]	[-6.69; -6.36]
$m = 31$	-8.15	[-2.86; -2.59]	[-5.92; -4.29]	[-7.23; -6.8]	[-8.16; -7.88]
$m = 36$	-9.07	NaN	[-6.5; -2.28]	[-8.24; -7.67]	[-9.01; -8.68]
$m = 41$	-9.65	NaN	[-7.94; -3.01]	[-9.64; -8.43]	[-10.35; -9.93]

(b) Given cost: $n = m$

Table 1: Approximation error ε in log-10 scale for the example 1. Abbreviations are defined in section 4.1.

4.3.2 A second function

In this section, we consider $\mathcal{X} = [-1, 1]$ equipped with the uniform measure and the function

$$u_2(x) = \frac{1}{1 + 5x^2}. \quad (35)$$

We consider the approximation space $V_m = \mathbb{P}_{m-1}$ with a basis $\{\varphi_i\}_{i=1}^m$ of Legendre polynomials. This is referred to as example 2. For this example, the same observations than in section 4.3.1 can be made:

- when resampling (see table 2a), it is possible to guarantee the stability of the approximation at a lower cost, without increasing the approximation error,
- when allowing the greedy algorithm to reach $n = m$, the stability of the **s-BLS** method is still verified for each of the 10 trials,
- the **s-BLS** method is comparable to interpolation in terms of the accuracy of the approximation (see table 2b).

The only difference is that the **EI** method behaves almost as well as the **I** method, which was not the case with the Gaussian measure.

4.3.3 A third function

We here consider the function

$$u_3(x) = \sum_{i=0}^p \exp(-\frac{i}{2}) \psi_i(x) \quad (36)$$

	OWLS		c-BLS ($M = 1$)		c-BLS ($M = 100$)	
	ε	n	ε	n	ε	n
$m = 6$	[-1.3; -1.3]	133	[-1.3; -1.25]	133	[-1.3; -1.28]	48
$m = 11$	[-2.43; -2.43]	265	[-2.43; -2.41]	265	[-2.42; -2.41]	108
$m = 16$	[-3.18; -3.18]	404	[-3.18; -3.16]	404	[-3.17; -3.16]	176
$m = 21$	[-4.31; -4.31]	548	[-4.31; -4.29]	548	[-4.31; -4.3]	249
$m = 26$	[-5.06; -5.06]	696	[-5.06; -5.05]	696	[-5.06; -5.04]	325
$m = 31$	[-6.19; -6.19]	847	[-6.19; -6.18]	847	[-6.19; -6.17]	405
$m = 36$	[-6.94; -6.94]	1001	[-6.94; -6.93]	1001	[-6.94; -6.93]	487
$m = 41$	[-8.07; -8.07]	1156	[-8.07; -8.07]	1156	[-8.07; -8.06]	571

(a) Guaranteed stability: the value of n is chosen to ensure a stable approximation with probability greater than 0.99.

	I	EI	OWLS	BLS ($M = 100$)	s-BLS
	ε	ε	ε	ε	ε
$m = 6$	-1.21	[-0.83; -0.83]	[-0.54; 0.82]	[-1.13; -0.88]	[-1.25; -1.12]
$m = 11$	-2.3	[-2.2; -2.2]	[-1.14; 1.29]	[-2.07; -1.5]	[-2.3; -2.11]
$m = 16$	-3.1	[-2.85; -2.85]	[-1.66; 2.6]	[-2.53; -2.17]	[-3.11; -2.75]
$m = 21$	-4.18	[-4.13; -3.96]	[-2.36; 1.05]	[-3.41; -2.49]	[-4.18; -4.03]
$m = 26$	-4.98	[-4.85; -4.85]	[-2; 3.25]	[-4.18; -2.96]	[-5.04; -4.75]
$m = 31$	-6.07	[-5.71; -5.71]	[-4.15; 2.94]	[-4.94; -3.78]	[-6.01; -5.92]
$m = 36$	-6.86	[-6.62; -6.48]	[-0.51; 3.45]	[-5.59; -4.2]	[-6.94; -6.69]
$m = 41$	-7.95	[-7.84; -7.61]	[-2.67; 1.91]	[-6.38; -5.24]	[-7.91; -7.72]

(b) Given cost: $n = m$.

Table 2: Approximation error ε in log-10 scale for the example 2. Abbreviations are defined in section 4.1.

where $\mathcal{X} = \mathbb{R}$ is equipped with the Gaussian measure, $(\psi_1, \dots, \psi_m) = \mathbf{U}(\varphi_1, \dots, \varphi_m)$, with $\{\varphi_i\}_{i=1}^m$ the set of Hermite polynomials of degree less than p and \mathbf{U} an orthogonal matrix. In practice \mathbf{U} is taken as the matrix of the left singular vectors of a $m \times m$ matrix \mathbf{A} , whose elements are i.i.d. realizations of a standard Gaussian random variable $\mathcal{N}(0, 1)$.

In this example, p is chosen equal to 40, the approximation space $V_m = \text{span}\{\psi_i : 1 \leq i \leq m\}$, and we consider different \mathbf{U} for each trial. Therefore, we also have confidence intervals for the **I** method. This is referred to as example 3 and the associated results are summarized in tables 3a and 3b. Hence, in the same manner than in tables 1 and 2, we notice that

- the error of approximation decreases when the size of the approximation space increases for all methods except **EI** (where the problem occurs when $m \geq 31$),
- the errors associated with **c-BLS** ($M = 100$) method are almost the same than the ones associated with the **OWLS** and **c-BLS** ($M = 1$) methods while being based on less evaluations of the function,
- the **s-BLS** method provides better results than the **OWLS** and **BLS** ($M = 100$) methods when n is chosen equal to m (interpolation regime).

For this example, it is important to notice that the approximation space is not generated from a set of commonly-used polynomials, for which there exists adapted sequences of points for interpolation. However for the **I** method, we still use Gauss-Hermite points, which may explain why the **s-BLS** method outperforms the **I** method in table 3b. This highlights the interest of the **s-BLS** method, which guarantees good sequences of points for the approximation,

	OWLS		c-BLS ($M = 1$)		c-BLS ($M = 100$)	
	ε	n	ε	n	ε	n
$m = 6$	[-1.22; -1.20]	133	[-1.42; -1.40]	133	[-1.41; -1.37]	48
$m = 11$	[-2.38; -2.30]	265	[-2.57; -2.47]	265	[-2.57; -2.48]	108
$m = 16$	[-3.44; -3.34]	404	[-3.62; -3.50]	404	[-3.61; -3.52]	176
$m = 21$	[-4.50; -4.42]	548	[-4.71; -4.63]	548	[-4.69; -4.60]	249
$m = 26$	[-5.62; -5.48]	696	[-5.86; -5.68]	696	[-5.82; -5.70]	325
$m = 31$	[-6.72; -6.59]	847	[-6.93; -6.81]	847	[-6.93; -6.78]	405
$m = 36$	[-7.73; -7.63]	1001	[-7.93; -7.83]	1001	[-7.94; -7.80]	487
$m = 41$	[-15.09; -14.63]	1156	[-15.35; -14.93]	1156	[-15.39; -15.02]	571

(a) Guaranteed stability: the value of n is chosen to ensure a stable approximation with probability greater than 0.99.

	I	EI	OWLS	BLS ($M = 100$)	s-BLS
	ε	ε	ε	ε	ε
$m = 6$	[-0.95; 0.0173]	[-1.27; -1.07]	[-1.02; 0.044]	[-1.31; -1.09]	[-1.14; -0.96]
$m = 11$	[-2.13; -1.46]	[-2.36; -2.05]	[-2.04; -0.99]	[-2.35; -2.07]	[-2.19; -2.06]
$m = 16$	[-3.07; -2.46]	[-3.25; -2.89]	[-3.04; -2.30]	[-3.28; -2.80]	[-3.29; -3.03]
$m = 21$	[-4.15; -3.61]	[-4.08; -3.69]	[-3.90; -3.29]	[-4.18; -3.89]	[-4.37; -4.11]
$m = 26$	[-5.01; -3.64]	[-5.11; -4.38]	[-4.97; -2.93]	[-5.15; -4.81]	[-5.48; -5.17]
$m = 31$	[-6.22; -5.33]	[-6.19; -5.05]	[-6.02; -5.67]	[-6.27; -5.80]	[-6.46; -6.31]
$m = 36$	[-7.14; -6.48]	[-3.79; -2.13]	[-7.02; -5.61]	[-7.39; -6.51]	[-7.58; -7.34]
$m = 41$	[-15.34; -15.28]	NaN	[-12.40; -5.27]	[-12.95; -12.32]	[-15.03; -14.86]

(b) Given cost: $n = m$

Table 3: Approximation error ε for the example 3. Abbreviations are defined in section 4.1.

no matter what the approximation space is. The results obtained with the **I** method may be improved by choosing a suitable set of initial points (in size and distribution).

Using the same function u_3 , but with \mathcal{X} equipped with the uniform measure μ and $\{\varphi_i\}_{i=1}^m$ the set of Legendre polynomials we draw the same conclusions than with the Gaussian measure, except that the **EI** method does not converge anymore to machine precision (due to numerical instabilities).

5 Conclusion

We have proposed a method to construct the projection of a function u in a given approximation space V_m with dimension m . In this method, the approximation is a weighted least-squares projection associated with random points sampled from a suitably chosen distribution. We obtained quasi-optimality properties (in expectation) for the weighted least-squares projection, with or without reducing the size of the sample by a greedy removal of points.

The error bound in the quasi-optimality property depends on the number of points selected by the greedy algorithm. The more points removed, the larger the bound will be. Therefore, if the goal is an accurate control of the error, as few points as possible should be removed. On the contrary, if the goal is to reduce the cost as much as possible but allows a larger bound of the error, the maximum number of points may be removed from the sample, which in some cases leads to an interpolation regime ($n = m$).

As the convergence of this greedy algorithm to the interpolation regime is not systematic, it would be interesting to look for an optimal selection of the sub-sample with regard to the sta-

bility criterion.

With this method, the points are sampled from a distribution which depends on the approximation space. Considering strategies where this approximation space is chosen adaptively, as in [1], an important issue is the reuse of samples from one approximation space to another. In this article, we have only considered the case of noiseless data, but an extension to noisy data could be considered.

A Proof of Lemma 3.4

Recall that for any sample \mathbf{x}^n , $Z_{\mathbf{x}^n} = \|\mathbf{G}_{\mathbf{x}^n} - \mathbf{I}\|_2$ and $\mathbb{P}(A_\delta) \geq 1 - \eta^M$ (Lemma 3.1). By definition of $\mathbf{x}^{n,\star}$, we have $\mathbf{x}^{n,\star} = \mathbf{x}^{n,I^\star}$, where given the M samples $\mathbf{x}^{n,1}, \dots, \mathbf{x}^{n,M}$, the random variable I^\star follows the uniform distribution on the set $\arg \min_{1 \leq i \leq M} Z_{\mathbf{x}^{n,i}}$ (possibly reduced to a singleton). The property (22) is a particular case of (21) for $\varepsilon = 1$. However, let us first provide a simple proof of (22). We have

$$\begin{aligned} \mathbb{E}(\|v\|_{\mathbf{x}^n}^2) &= \mathbb{E}(\|v\|_{\mathbf{x}^{n,\star}}^2 | A_\delta) \leq \mathbb{E}(\|v\|_{\mathbf{x}^{n,\star}}^2) \mathbb{P}(A_\delta)^{-1} \\ &\leq \mathbb{E}(\|v\|_{\mathbf{x}^{n,I^\star}}^2) (1 - \eta^M)^{-1} \leq \sum_{j=1}^M \mathbb{E}(\|v\|_{\mathbf{x}^{n,j}}^2) (1 - \eta^M)^{-1} \\ &= \|v\|^2 M (1 - \eta^M)^{-1}. \end{aligned}$$

Now let us consider the proof of the other inequalities. We first note that $A_\delta = \{Z_{\mathbf{x}^{n,I^\star}} \leq \delta\} = \{\min_{1 \leq i \leq M} Z_{\mathbf{x}^{n,i}} \leq \delta\}$. We consider the events $B_j = \{I^\star = j\}$ which form a complete set of events. From the definition of I^\star and A_δ and the fact that the samples $\mathbf{x}^{n,i}$ are i.i.d., it is clear that $\mathbb{P}(B_j) = \mathbb{P}(B_1) = M^{-1}$ and $\mathbb{P}(B_j \cap A_\delta) = \mathbb{P}(B_1 \cap A_\delta)$ for all j . Therefore,

$$\mathbb{P}(A_\delta \cap B_1) = \frac{1}{M} \sum_{j=1}^M \mathbb{P}(A_\delta \cap B_j) = \frac{1}{M} \mathbb{P}(A_\delta) \geq \frac{(1 - \eta^M)}{M}.$$

Then

$$\begin{aligned} \mathbb{E}(\|v\|_{\mathbf{x}^n}^2) &= \mathbb{E}(\|v\|_{\mathbf{x}^{n,\star}}^2 | A_\delta) = \sum_{j=1}^M \mathbb{E}(\|v\|_{\mathbf{x}^{n,j}}^2 | A_\delta \cap B_j) \mathbb{P}(B_j) = \mathbb{E}(\|v\|_{\mathbf{x}^{n,1}}^2 | A_\delta \cap B_1) \\ &= \mathbb{E}(\|v\|_{\mathbf{x}^{n,1}}^2 \mathbb{1}_{A_\delta \cap B_1}) \mathbb{P}(A_\delta \cap B_1)^{-1} \\ &\leq \mathbb{E}(\|v\|_{\mathbf{x}^{n,1}}^2 \mathbb{1}_{Z_{\mathbf{x}^{n,1}} \leq \delta} \mathbb{1}_{\min_{2 \leq i \leq M} Z_{\mathbf{x}^{n,i}} \geq Z_{\mathbf{x}^{n,1}}}) M (1 - \eta^M)^{-1} \\ &= \mathbb{E}(\|v\|_{\mathbf{x}^{n,1}}^2 \mathbb{1}_{Z_{\mathbf{x}^{n,1}} \leq \delta} \mathbb{E}(\mathbb{1}_{\min_{2 \leq i \leq M} Z_{\mathbf{x}^{n,i}} \geq Z_{\mathbf{x}^{n,1}}} | \mathbf{x}^{n,1})) M (1 - \eta^M)^{-1} \\ &= \mathbb{E}(\|v\|_{\mathbf{x}^{n,1}}^2 \mathbb{1}_{Z_{\mathbf{x}^{n,1}} \leq \delta} \mathbb{E}(\mathbb{1}_{Z_{\mathbf{x}^{n,2}} > Z_{\mathbf{x}^{n,1}}} | \mathbf{x}^{n,1})^{M-1}) M (1 - \eta^M)^{-1}. \end{aligned}$$

Using Hölder's inequality, we have that for any $0 < \varepsilon \leq 1$,

$$\begin{aligned} \mathbb{E}(\|v\|_{\mathbf{x}^{n,\star}}^2 | A_\delta) &\leq \mathbb{E}\left(\|v\|_{\mathbf{x}^{n,1}}^{\frac{2}{\varepsilon}} \mathbb{1}_{Z_{\mathbf{x}^{n,1}} \leq \delta}\right)^\varepsilon \mathbb{E}\left(\mathbb{E}\left(\mathbb{1}_{Z_{\mathbf{x}^{n,2}} > Z_{\mathbf{x}^{n,1}}} | \mathbf{x}^{n,1}\right)^{\frac{M-1}{1-\varepsilon}}\right)^{1-\varepsilon} M (1 - \eta^M)^{-1} \\ &\leq \mathbb{E}\left(\|v\|_{\mathbf{x}^{n,1}}^{\frac{2}{\varepsilon}}\right)^\varepsilon \mathbb{E}\left(\mathbb{E}\left(\mathbb{1}_{Z_{\mathbf{x}^{n,2}} > Z_{\mathbf{x}^{n,1}}} | \mathbf{x}^{n,1}\right)^{\frac{M-1}{1-\varepsilon}}\right)^{1-\varepsilon} M (1 - \eta^M)^{-1}. \end{aligned}$$

For any measurable function f and any two i.i.d. random variables X and Y , we have that $\mathbb{E}(\mathbb{1}_{f(X) > f(Y)} | Y)$ is a uniform random variable on $(0, 1)$. Therefore $\mathbb{E}(\mathbb{1}_{Z_{\mathbf{x}^{n,2}} > Z_{\mathbf{x}^{n,1}}} | \mathbf{x}^{n,1})$ is uniformly distributed on $(0, 1)$ and

$$\mathbb{E}\left(\mathbb{E}\left(\mathbb{1}_{Z_{\mathbf{x}^{n,2}} > Z_{\mathbf{x}^{n,1}}} | \mathbf{x}^{n,1}\right)^{\frac{M-1}{1-\varepsilon}}\right) = \frac{1}{\frac{M-1}{1-\varepsilon} + 1} = \frac{1 - \varepsilon}{M - \varepsilon}.$$

By combining the previous results, we obtain

$$\mathbb{E}(\|v\|_{\mathbf{x}^{n,*}}^2 | A_\delta) \leq \mathbb{E}(\|v\|_{\mathbf{x}^n}^{\frac{2}{\varepsilon}})^\varepsilon M(1 - \eta^M)^{-1} \frac{(1 - \varepsilon)^{1-\varepsilon}}{(M - \varepsilon)^{1-\varepsilon}}.$$

For $\varepsilon = 1$, we recover the result (22). The last result simply follows from

$$\mathbb{E}(\|v\|_{\mathbf{x}^n}^{\frac{2}{\varepsilon}}) \leq \mathbb{E}(\|v\|_{\mathbf{x}^n}^2) \|v\|_{\infty,w}^{2/\varepsilon-2} = \|v\|^2 \|v\|_{\infty,w}^{2/\varepsilon-2}.$$

References

- [1] B. Arras and Cohen A. Bachmayr M. Sequential sampling for optimal weighted least squares approximations in hierarchical spaces. *arXiv:1805.10801*, 12, 2018.
- [2] A. Cohen and G. Migliorati. Optimal weighted least-squares methods. *SMAI Journal of Computational Mathematics*, 3:181–203, 2017.
- [3] M.A. Cohen, A. Davenport and D. Leviatan. On the stability and accuracy of least squares approximation. *Foundations of Computational Mathematics*, 13:819–834, 2013.
- [4] A. Doostan and J. Hampton. Coherence motivated sampling and convergence analysis of least squares polynomial chaos regression. *Computer Methods in Applied Mechanics and Engineering*, 290:73–97, 2015.
- [5] Y. Maday, N.C. Nguyen, A. Patera, and G. Pau. A general multipurpose interpolation procedure: the magic points. *Communications on Pure and Applied Analysis*, 8(1):383–404, 2009.
- [6] Akil Narayan, John Jakeman, and Tao Zhou. A christoffel function weighted least squares algorithm for collocation approximations. *Mathematics of Computation*, 86, 05 2017.
- [7] A.J. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12:389–434, 2012.