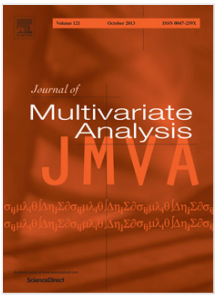
Accepted Manuscript



Convergence estimates in probability and in expectation for discrete least squares with noisy evaluations at random points

Giovanni Migliorati, Fabio Nobile, Raúl Tempone

PII: S0047-259X(15)00193-1
DOI: http://dx.doi.org/10.1016/j.jmva.2015.08.009
Reference: YJMVA 3983

To appear in: *Journal of Multivariate Analysis*

Received date: 24 December 2014

Please cite this article as: G. Migliorati, F. Nobile, R. Tempone, Convergence estimates in probability and in expectation for discrete least squares with noisy evaluations at random points, *Journal of Multivariate Analysis* (2015), http://dx.doi.org/10.1016/j.jmva.2015.08.009

Convergence estimates in probability and in expectation for discrete least squares with noisy evaluations at random points

Giovanni Migliorati* Fabio Nobile† Raúl Tempone‡

July 10, 2015

Abstract

We study the accuracy of the discrete least-squares approximation on a finite dimensional space of a real-valued target function from noisy pointwise evaluations at independent random points distributed according to a given sampling probability measure. The convergence estimates are given in mean-square sense with respect to the sampling measure. The noise may be correlated with the location of the evaluation and may have nonzero mean (offset). We consider both cases of bounded or square-integrable noise / offset. We prove conditions between the number of sampling points and the dimension of the underlying approximation space that ensure a stable and accurate approximation. Particular focus is on deriving estimates in probability within a given confidence level. We analyze how the best approximation error and the noise terms affect the convergence rate and the overall confidence level achieved by the convergence estimate. The proofs of our convergence estimates in probability use arguments from the theory of large deviations to bound the noise term. Finally we address the particular case of multivariate polynomial approximation spaces with any density in the beta family, including uniform and Chebyshev.

Keywords: approximation theory, discrete least squares, noisy evaluations, error analysis, convergence rates, large deviations, learning theory, multivariate polynomial approximation.

MSC: 41A10, 41A25, 41A50, 41A63, 62G08, 65M70.

1 Introduction

The motivations of our analysis come from the development of discrete least-squares approximation methods for functions depending on a multivariate random variable distributed according to a known probability measure. This topic falls at the intersection of approximation theory and learning theory [6, 7], and is related to nonparametric regression with random design [10] and statistical learning theory [22]. More specifically, our framework is an instance of the *projection learning problem* (or improper function learning problem) described in [6, 18, 19].

We focus on the discrete least-squares approximation of a target function on a given finite dimensional (linear) vector space using pointwise evaluations at independent and randomly selected

*MATHICSE-CSQI, École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland. email: giovanni.migliorati@epfl.ch

†MATHICSE-CSQI, École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland. email: fabio.nobile@epfl.ch

‡Applied Mathematics and Computational Sciences, and SRI Center for Uncertainty Quantification in Computational Science and Engineering, KAUST, Thuwal 23955-6900, Saudi Arabia. email: raul.tempone@kaust.edu.sa

points, identically distributed according to the underlying probability measure. In particular, we are interested in the case of discrete least-squares projection on not necessarily bounded multivariate approximation sets, *i.e.* the minimizer of the discrete least-squares problem is not constrained to be in a compact subset. Two situations might occur, depending on the context and on the origin of the evaluations of the target function: *noiseless* evaluations or *noisy* evaluations. The former situation arises for example in an abstract modeling context, where round-off or other discretization errors can be properly controlled. The latter situation typically arises when dealing with experimental data, which are polluted by measurement and/or systematic errors.

A vast literature is available for discrete least-squares approximations on compact sets or linear vector spaces in the noisy case. In the case of linear vector spaces, we mention the bound in [10, Theorem 11.3] or those in [2, 3], which hold in expectation under the assumption that the target function itself is bounded. Often a truncation operator has to be used to obtain those bounds. Moreover, these results are nonoptimal in the noiseless case, as the best approximation error in the subspace is not recovered when the amount of noise tends to zero.

The stability and accuracy of discrete least squares on finite dimensional vector spaces in the noiseless and noisy cases have been recently analyzed in several works [5, 15, 4, 16, 12]. It is shown that optimal convergence rates can be recovered in the noiseless case if a suitable relation between the number of evaluations and the dimension of the approximation space is enforced. Moreover, such relation guarantees stability of the discrete projection with high probability.

Generalizations of the previous analyses to the noisy case have been presented as well in the aforementioned works. In the particular case of bounded noise (stochastic or deterministic) with zero mean, an estimate in expectation has been proposed in [5]. Estimates in expectation with the deterministic noise model have been proven in [4]. Estimates in probability have been proven in [4] but using the best approximation error in L^∞ rather than L^2 and focusing only on the deterministic noise model. In both the noiseless and noisy cases, the analyses in [5, 4] rely on the Chernoff bounds for sums of random matrices proven in [1, 20]. The analysis in [15] uses different techniques to derive a convergence estimate in probability, and covers only the noiseless case.

The purpose of the present work is to derive new convergence estimates in probability and in expectation, in the general case of noise of stochastic type with nonzero mean, that recover optimal convergence rates in the limit of zero noise. We split the noise into two parts: the conditional expectation of the noise w.r.t. the sampling measure, that we name in the following as the *offset* of the noise, and the part of the noise due to its intrinsic randomness, hereafter called fluctuations. According to this splitting, we consider three types of noise models: (i) square-integrable offset and uniformly bounded conditional variance of the fluctuations with respect to the sampling measure, (ii) square-integrable offset and bounded fluctuations, (iii) bounded offset and fluctuations. Using arguments coming from the theory of large deviations [8, 21], we prove in Theorem 9 a probabilistic bound for the fluctuation term in the discrete least-square projection, *i.e.* taking out the effect of the offset. Afterwards, exploiting Theorem 9, for each one of the aforementioned noise models we prove convergence estimates in probability for the discrete least-square projection error when a specific condition is satisfied between the number of pointwise evaluations and the dimension of the underlying approximation space. The derived convergence estimates relate the L^2 approximation error of the discrete least-squares approximation with the best approximation error measured either in the L^2 norm or in the L^∞ norm. These probability estimates do not require the use of any truncation operator. Moreover, we prove a convergence estimate in expectation with the unbounded noise model, that generalizes a result previously given in [5] to the case of nonzero offset. Our convergence estimates, both in probability and in expectation, separate the contribution to the error due to the best approximation error on a given approximation space and the contribution

due to the presence of noise, similarly to the so-called *bias-variance trade off*, see *e.g.* [6, 17].

Finally we apply our results to the particular setting of multivariate polynomial approximation spaces, which is a provably effective choice in many situations where a smooth dependence on many parameters needs to be approximated. Examples of such a situation arise when approximating the parameter-to-solution map of many types of PDEs with stochastic data, see *e.g.* the monographs [9, 11] or the works [4, 12, 16] focused on discrete least squares. In [5, 15, 4], discrete least squares on multivariate polynomial spaces with evaluations at random points have been analyzed with the uniform and arcsine density: in any dimension and with polynomial spaces associated with downward closed multi-index sets, stability and accuracy have been proven, provided a specific proportionality relation is satisfied between the number of evaluations and the dimension of the polynomial approximation space. Then the analysis has been extended to any density in the beta family, using the results proven in [13].

In [14] it has been proven that, in the case of uniform density and with anisotropic tensor product polynomial spaces in any dimension, the random point set can be replaced by suitable low-discrepancy point sets, leading to analogous results concerning stability and accuracy of discrete least squares in the noiseless case. These results can be combined with those of the present paper, to provide convergence estimates for discrete least squares with noisy evaluations at low-discrepancy point sets, rather than random point sets.

Another analysis of discrete least squares with deterministic points has been proposed in [23], with points that are asymptotically distributed according to the arcsine density.

The outline of the paper is the following. In Section 2 we introduce the discrete least-squares approximation, the observation models, the assumptions on the noise and the noise models. In Section 2.1 we briefly present the algebraic formulation of discrete least squares and in Section 2.2 we recall the results achieved in [5]. In Section 3 we present our estimates in expectation (Section 3.1) and in probability (Section 3.2). Several intermediate results used in the proofs of these estimates have been collected in Section 5 where, in particular, we derive an estimate for the noise term using arguments from the theory of large deviations. In Section 4 we apply our convergence estimates in the noisy case to the setting of polynomial approximation. Finally in Section 6 we draw some conclusions.

2 Discrete least squares with noisy evaluations at random points

Let $\Gamma \subseteq \mathbb{R}^d$ be a subset of the d-dimensional Euclidean space such that $\Gamma = \prod_{i=1}^d \Gamma_i$, with $\Gamma_i \subseteq \mathbb{R}$ being closed intervals for any $i = 1, \dots, d$ and $N \subseteq \mathbb{R}$ a subset of $\mathbb{R}$. We introduce a complete probability space (Ω, Σ, ν), with $\Omega := \Gamma \times N$ being the sample space, Σ the σ-algebra of Borel sets and ν a probability measure. For a random variable $(y, \eta) \in \Gamma \times N$ distributed according to the joint measure ν, we denote by μ the marginal probability measure with respect to y, *i.e.* $\mu(B) = \nu(B \times N)$ for any Borel set $B \in \Gamma$. Moreover, we assume that μ is absolutely continuous with respect to the Lebesgue measure λ on Γ and denote with $\rho : \Gamma \to \mathbb{R}^+$, $\rho = d\mu/d\lambda$ the associated probability density function.

We introduce a given target function $u : \Gamma \to \mathbb{R}$, that we would like to approximate in the L^2-probability sense using pointwise evaluations $u(y_1), \dots, u(y_m)$ in m independent and randomly chosen points $y_1, \dots, y_m \in \Gamma$ distributed according to the measure μ. We assume that the function u is well-defined at any point in Γ except eventually a zero μ-measure set and that $u \in L^2_\mu := \{v : \Gamma \to \mathbb{R} : \int_\Gamma v^2 d\mu < +\infty\}$. Hereafter, the L^2_μ norm will be denoted simply by $\|\cdot\|$, *i.e.*

$\|v\|_{L^2_\mu(\Gamma)} = \|v\| = (\int_\Gamma v^2 d\mu)^{1/2}$.

The evaluations $u(y_1), \dots, u(y_m)$ are eventually polluted by noise, coming from any source of uncertainty due to controlled or uncontrolled agents. In the present paper we consider a *stochastic noise model*, where the noise is described by means of random variables. We define the noiseless and noisy observation models as

$$\textbf{noiseless model}, \quad z_j := u(y_j), \quad j = 1, \dots, m, \tag{1}$$

$$\textbf{noisy model}, \quad z_j := u(y_j) + \eta_j, \quad j = 1, \dots, m, \tag{2}$$

where $(y_1, \eta_1), \dots, (y_m, \eta_m) \in \Omega$ are m i.i.d. random variables distributed according to the joint probability measure ν. When collecting the measurements $z_1, \dots, z_m$, we sample the random variable (y, η) and observe the couple (y_j, η_j) for any $j = 1, \dots, m$: we have independence among the realizations of the couple, but in general the random variables η and y are mutually dependent. Of course the noiseless case can be seen as a particular instance of the noisy case with $\eta_j = 0$ for any $j = 1, \dots, m$.

For any $m \geq 1$, we define also the sample product space Ω^m, the σ-algebra Σ^m and the product measure ν^m as

$$\Omega^m := \underbrace{\Omega \times \cdots \times \Omega}_{m \text{ times}}, \quad \Sigma^m := \prod_{i=1}^m \Sigma_i, \quad \nu^m := \underbrace{\nu \otimes \cdots \otimes \nu}_{m \text{ times}}.$$

We can finally introduce the complete probability space $(\Omega^m, \Sigma^m, \nu^m)$ that characterizes the settings of our discrete least-squares approximation using m pointwise noisy evaluations. Unless mentioned otherwise, throughout the paper Pr and $\mathbb{E}$ refer to the probability and the expectation w.r.t. the joint measure ν. Moreover, we define the sets Γ^m and N^m as

$$\Gamma^m := \underbrace{\Gamma \times \cdots \times \Gamma}_{m \text{ times}}, \quad N^m := \underbrace{N \times \cdots \times N}_{m \text{ times}}.$$

We define the conditional mean of the noise as

$$\overline{\eta}(y) := \mathbb{E}(\eta | y), \tag{3}$$

and sometimes use the more concise term *noise offset* to refer to (3). The (total) expectation of the random variable $\overline{\eta}^2$ is given by

$$\mathbb{E}\left(\overline{\eta}^2\right) = \int_\Gamma \int_{\mathbb{R}} \overline{\eta}(y)^2 d\nu(y, \eta) = \int_\Gamma \overline{\eta}(y)^2 d\mu(y) = \|\overline{\eta}\|^2.$$

We define the inner product

$$\langle f_1, f_2 \rangle := \int_\Gamma f_1(y) f_2(y) d\mu(y), \quad \forall f_1, f_2 \in L^2_\mu(\Gamma), \tag{4}$$

as well as the discrete inner product

$$\langle f_1, f_2 \rangle_m := m^{-1} \sum_{j=1}^m f_1(y_j) f_2(y_j), \quad \forall f_1, f_2 \in L^2_\mu(\Gamma),$$

with $y_1, \dots, y_m$ being independent random points in Γ identically distributed according to the measure μ. These inner products are associated with the norm $\|\cdot\| = \langle \cdot, \cdot \rangle^{1/2}$ (already previously defined) and seminorm $\|\cdot\|_m := \langle \cdot, \cdot \rangle_m^{1/2}$, respectively. Notice that $\mathbb{E}(\|\cdot\|_m) = \|\cdot\|$. We denote by $V_n \subset L^\infty(\Gamma)$ any finite dimensional subspace of $L^2_\mu(\Gamma)$ such that $n := \dim(V_n)$, and by $(\psi_i)_{1 \leq i \leq n}$ an orthonormal basis with the canonical inner product (4). Concrete examples of finite dimensional spaces of multivariate polynomials are given in Section 4.

Given the target function $u : \Gamma \to \mathbb{R}$, we define its continuous L^2 projection over V_n as

$$\Pi_n u := \operatorname*{argmin}_{v \in V_n} \|u - v\|,$$

and denote by

$$e_n(u) := \inf_{v \in V_n} \|u - v\| = \|u - \Pi_n u\|,$$

its best approximation error in the L^2_μ norm. We denote by

$$e_n^\infty(u) := \inf_{v \in V_n} \|u - v\|_{L^\infty}$$

the best approximation error in the L^∞ norm. We also define the *discrete L^2 projection* over V_n of the noiseless or noisy evaluations of u in the points $y_1, \dots, y_m$ as

$$w := \operatorname*{argmin}_{v \in V_n} \sum_{i=1}^{m} |z_i - v(y_i)|^2 . \tag{5}$$

This minimization corresponds to minimize the discrete seminorm containing the evaluations (eventually polluted by noise) of the target function u in the m points $y_1, \dots, y_m \in \Gamma$. In the noiseless case, we replace the notation w with $\Pi_n^m u$ to emphasize the lack of noise. Given a threshold $\tau \in \mathbb{R}_0^+$, we introduce the truncation operator

$$T_\tau(t) := \operatorname{sign}(t) \min\{\tau, |t|\}, \quad \text{for any } t \in \mathbb{R},$$

and use it to define the truncated discrete L^2 projection in the noiseless or noisy cases, respectively as:

$$\widetilde{\Pi}_n^m := T_\tau \circ \Pi_n^m, \quad \text{and} \quad \widetilde{w} := T_\tau \circ w.$$

It is convenient to introduce the notation $\mathbf{y} \in \Gamma^m$ to denote the "vector" containing the m points $y_1, \dots, y_m \in \Gamma$. We denote with $\mathbf{z} \in \mathbb{R}^m$ the vector containing the observations $z_1, \dots, z_m$. The vector $\mathbf{u} \in \mathbb{R}^m$ contains the evaluations of the target function $u(y_1), \dots, u(y_m)$ in the points, and the vector $\boldsymbol{\eta} \in \mathbb{R}^m$ contains the noise $\eta_1, \dots, \eta_m$ so that

$$\mathbf{z} = \mathbf{u} + \boldsymbol{\eta}.$$

We also introduce the vectors $\mathbf{g}, \overline{\boldsymbol{\eta}}, \widetilde{\boldsymbol{\eta}} \in \mathbb{R}^m$ with elements given by

$$[\mathbf{g}]_j := u(y_j) - \Pi_n u(y_j), \quad j = 1, \dots, m, \tag{6}$$

$$[\overline{\boldsymbol{\eta}}]_j := \overline{\eta}(y_j), \quad j = 1, \dots, m, \tag{7}$$

$$[\widetilde{\boldsymbol{\eta}}]_j := \eta_j - \overline{\eta}(y_j), \quad j = 1, \dots, m. \tag{8}$$

Hence, the following splitting for the noise vector $\boldsymbol{\eta}$ holds true:

$$\boldsymbol{\eta} = \overline{\boldsymbol{\eta}} + \widetilde{\boldsymbol{\eta}}. \tag{9}$$

We define the supremum of the noise terms $\widetilde{\eta} = \widetilde{\eta}(y)$ and $\overline{\eta} = \overline{\eta}(y)$ as

$$\widetilde{\eta}_{\max} := \sup_{y \in \Gamma} |\widetilde{\eta}(y)| \qquad \text{and} \qquad \overline{\eta}_{\max} := \sup_{y \in \Gamma} |\overline{\eta}(y)|.$$

We also define

$$\eta_{\max} := \widetilde{\eta}_{\max} + \overline{\eta}_{\max} < +\infty.$$

Depending on the modeling context, the noise offset might be square-integrable

$$\|\overline{\eta}\| < +\infty, \tag{10}$$

and/or the noise might have uniformly bounded conditional variance

$$\sigma^2 := \max_{y\in\Gamma} \mathbb{E}(|\eta - \overline{\eta}(y)|^2|y) < +\infty. \tag{11}$$

We now precisely define several noise models, eventually making use of the previous conditions (10)–(11) on the noise. Three different situations will be addressed:

$$\textbf{bounded noise}, \begin{cases} |\widetilde{\eta}_{\max}| < +\infty,\ \mathrm{Im}(\widetilde{\eta}_j) \subseteq [-\widetilde{\eta}_{\max}, \widetilde{\eta}_{\max}],\ \forall j = 1,\dots,m, \\ |\overline{\eta}_{\max}| < +\infty,\ \mathrm{Im}(\overline{\eta}_j) \subseteq [-\overline{\eta}_{\max}, \overline{\eta}_{\max}],\ \forall j = 1,\dots,m, \end{cases} \tag{12}$$

$$\textbf{unbounded offset and bounded fluctuations}, \begin{cases} |\widetilde{\eta}_{\max}| < +\infty,\ \mathrm{Im}(\widetilde{\eta}_j) \subseteq [-\widetilde{\eta}_{\max}, \widetilde{\eta}_{\max}],\ \forall j = 1,\dots,m, \\ \mathrm{Im}(\overline{\eta}_j) = \mathbb{R},\ \forall j = 1,\dots,m, \\ \|\overline{\eta}\| < +\infty, \end{cases} \tag{13}$$

$$\textbf{unbounded noise}, \begin{cases} \mathrm{Im}(\eta_j) = \mathbb{R},\ \forall j = 1,\dots,m, \\ \|\overline{\eta}\| < +\infty, \\ \sigma^2 < +\infty. \end{cases} \tag{14}$$

In the case of bounded noise, both the random variables $\widetilde{\eta}$ and $\overline{\eta}$ are bounded. This case contains for instance probability distributions in the beta family. The case of unbounded offset allows the offset random variable $\overline{\eta}$ to be unbounded and square-integrable. In the case of unbounded noise, both the random variables $\widetilde{\eta}$ and $\overline{\eta}$ are unbounded, the offset is square-integrable and the noise has uniformly bounded conditional variance. This case includes the Gaussian distribution, Laplace distribution and others.

2.1 Algebraic formulation of discrete least squares

We introduce the design matrix $\mathbf{D} \in \mathbb{R}^{m\times n}$, defined element-wise as $\mathbf{D}_{jk} := \psi_k(y_j)$ for any $j = 1,\dots,m$ and any $k = 1,\dots,n$, the Gramian matrix $\mathbf{G} := m^{-1}\mathbf{D}^\top\mathbf{D} \in \mathbb{R}^{n\times n}$, the matrix $\mathbf{J} := m^{-1}\mathbf{D} \in \mathbb{R}^{m\times n}$ and denote with $\mathbf{I}$ the $n \times n$ identity matrix. The matrix $\mathbf{D} = \mathbf{D}(\mathbf{y})$ depends on the points $y_1,\dots,y_m$ and is therefore random, and the matrices $\mathbf{G} = \mathbf{G}(\mathbf{y})$ and $\mathbf{J} = \mathbf{J}(\mathbf{y})$ depend on $\mathbf{y}$ through $\mathbf{D}$, but sometimes we omit to indicate the dependence on $\mathbf{y}$ that would unnecessarily overload the notation.

The discrete least-squares projection (5) can be calculated by solving the normal equations

$$\mathbf{G}w = \mathbf{J}^\top\mathbf{z}. \tag{15}$$

The right-hand side in (15) can be written as

$$\mathbf{J}^\top\mathbf{z} = \mathbf{J}^\top\mathbf{u} + \mathbf{J}^\top\boldsymbol{\eta} = \mathbf{J}^\top\mathbf{u} + \mathbf{J}^\top\overline{\boldsymbol{\eta}} + \mathbf{J}^\top\widetilde{\boldsymbol{\eta}},$$

separating the contribution due to the noise term and using the splitting (9).

From the definition of Gramian matrix it holds that

$$\mathrm{tr}(\mathbf{G}) = \sum_{k=1}^{n} \|\psi_k\|_m^2, \tag{16}$$

and

$$\mathbb{E}[\mathrm{tr}(\mathbf{G}(\mathbf{y}))] = n. \tag{17}$$

2.2 Previous analyses of discrete least squares

In this section we briefly summarize the results achieved in previous analyses of the stability and accuracy of discrete least squares. Following [4], we introduce the finite quantity

$$K(V_n) := \sup_{v \in V_n \setminus \{v \equiv 0\}} \frac{\|v\|_{L^\infty}^2}{\|v\|^2} = \sup_{y \in \Gamma} \sum_{k=1}^n |\psi_k(y)|^2 \leq \sum_{k=1}^n \|\psi_k\|_{L^\infty(\Gamma)}^2 < +\infty \tag{18}$$

that does not depend on the particular choice of the orthonormal basis, but only depends on V_n and ρ. In the case of multivariate polynomial orthonormal basis in any dimension, this quantity has been studied in [4] for Legendre and Chebyshev polynomials of the second type, and in [13] in the general case of Jacobi polynomials. We postpone the results obtained for polynomial approximation to Section 4.

In the following $|||\cdot|||$ always denotes the spectral matrix norm, and $\|\mathbf{v}\|_{\ell^2}$ denotes the Euclidean norm of any vector $\mathbf{v} \in \mathbb{R}^n$.

For any $\delta \in (0,1)$, we define $\zeta(\delta) := \delta + (1-\delta)\ln(1-\delta) > 0$. For any $r > 0$, consider now the following condition between the number of points m and the quantity $K(V_n)$:

$$\frac{m}{\ln m} \geq \frac{K(V_n)}{\kappa(\delta)}, \quad \kappa(\delta) := \frac{\zeta(\delta)}{1+r}. \tag{19}$$

For any $r > 0$ and any $\delta \in (0,1)$ define

$$\epsilon(m, \delta, r) := \frac{4\zeta(\delta)}{(1+r)\ln(m)} \leq \frac{4\zeta(\delta)}{\ln(m)}.$$

Notice that $\epsilon = \epsilon(m, \delta, r)$ is a decreasing function of m and r. We postpone to a few line below further information on the role of the parameter δ.

The main results in [5] imply that, for any $r > 0$ and a number of samples m large enough such that (19) is fulfilled, the following holds:

- the deviation between $\mathbf{G}$ and $\mathbf{I}$ satisfies

$$\Pr\{|||\mathbf{G} - \mathbf{I}||| > \delta\} \leq 2m \exp\left\{-\frac{\zeta(\delta) m}{K(V_n)}\right\}, \tag{20}$$

- in the noiseless case, if u satisfies a uniform bound τ over Γ, *i.e.* $|u(y)| \leq \tau$ a.s. w.r.t. μ, then

$$\mathbb{E}(\|u - \widetilde{\Pi}_n^m u\|^2) \leq (1 + \epsilon(m, \delta, r)) e_n(u)^2 + 8\tau^2 m^{-r}, \tag{21}$$

- in the noisy case, if u satisfies a uniform bound τ over Γ, *i.e.* $|u(y)| \leq \tau$ a.s. w.r.t. μ, then

$$\mathbb{E}(\|u - \widetilde{w}\|^2) \leq (1 + 2\epsilon(m, \delta, r)) e_n(u)^2 + 2(1-\delta)^{-2}\sigma^2 \frac{n}{m} + 8\tau^2 m^{-r}. \tag{22}$$

Moreover, the following events in the sample space Ω^m are equivalent

$$(1-\delta)\|v\|^2 \leq \|v\|_m^2 \leq (1+\delta)\|v\|^2, \quad \forall v \in V_n, \tag{23}$$

$$|||\mathbf{G} - \mathbf{I}||| \leq \delta, \tag{24}$$

$$1 - \delta \leq |||\mathbf{G}||| \leq 1 + \delta, \tag{25}$$

$$(1+\delta)^{-1} \leq |||\mathbf{G}^{-1}||| \leq (1-\delta)^{-1}, \tag{26}$$

$$m^{-1/2}(1-\delta)^{1/2} \leq |||\mathbf{J}||| \leq m^{-1/2}(1+\delta)^{1/2}, \tag{27}$$

and are subsets of the following larger event, belonging to Ω^m as well:

$$\mathrm{tr}(\mathbf{G}) \leq (1+\delta)n. \tag{28}$$

Therefore, under condition (19) the events (23)–(28) hold true with overwhelming probability, *i.e.*

$$\begin{aligned}
\Pr\left(\mathrm{tr}(\mathbf{G}) \leq (1+\delta)n\right) &\geq \Pr\left((1-\delta)\|v\|^2 \leq \|v\|_m^2 \leq (1+\delta)\|v\|^2,\ \forall v \in V_n\right) \\
&= \Pr\left(|||\mathbf{G}-\mathbf{I}||| \leq \delta\right) \\
&= \Pr\left(1-\delta \leq |||\mathbf{G}||| \leq 1+\delta\right) \\
&= \Pr\left((1+\delta)^{-1} \leq |||\mathbf{G}^{-1}||| \leq (1-\delta)^{-1}\right) \\
&= \Pr\left(m^{-1/2}(1-\delta)^{1/2} \leq |||\mathbf{J}||| \leq m^{-1/2}(1+\delta)^{1/2}\right) \\
&\geq 1 - 2m\exp\left\{-\frac{\zeta(\delta)m}{K(V_n)}\right\}.
\end{aligned} \tag{29}$$

For convenience we introduce the event

$$\Omega_\delta^m := \{(\mathbf{y},\boldsymbol{\eta}) \in \Omega^m \ :\ |||\mathbf{G}(\mathbf{y})-\mathbf{I}||| \leq \delta\} \in \Sigma^m, \tag{30}$$

for which, under condition (19), we have from (20) that

$$\Pr(\Omega_\delta^m) \geq 1 - 2m^{-r}. \tag{31}$$

Moreover, we shall use in the following the short hand notation

$$\mathbb{E}_+(X) := \int_{\Omega_\delta^m} X d\nu^m$$

for any Σ^m-measurable random variable X.

3 Convergence estimates for discrete least squares

In this section we present estimates in expectation and in probability for the noisy observation model (2), with the noise models (12), (13) and (14). The results in expectation are collected in Section 3.1 and mostly recall earlier results given in [5]. The new results in probability are collected in Section 3.2. Notice that all these results hold true in the general case of noise with nonzero mean.

3.1 Convergence estimates in expectation

The following theorem generalizes to the case $\overline{\eta}(y) \neq 0$ the theorem given in [5, Theorem 3].

Theorem 1. *For any $r > 0$ and in the case of the noise model* (14) *with nonzero mean: if m satisfies condition* (19) *and u satisfies a uniform bound τ over Γ, then*

$$\mathbb{E}(\|u-\widetilde{w}\|^2) \leq \left(1 + \frac{2\zeta(\delta)}{(1-\delta)^2(1+r)\ln m}\right) e_n(u)^2 + \frac{2}{(1-\delta)^2}\left(\frac{\sigma^2 n}{m} + \|\overline{\eta}\|^2\left(1 + \frac{\zeta(\delta)}{(1+r)\ln m}\right)\right) + 8\tau^2 m^{-r}. \tag{32}$$

Proof. The proof follows closely the one given in [5, Theorem 3]. We have

$$\mathbb{E}(\|u-\widetilde{w}\|^2) = \int_{\Omega_\delta^m} \|u-\widetilde{w}\|^2 d\nu^m + \int_{\Omega^m \setminus \Omega_\delta^m} \|u-\widetilde{w}\|^2 d\nu^m \leq \mathbb{E}_+(\|u-w\|^2) + 8\tau^2 m^{-r}.$$

Using the results in Lemma 2 in Section 5, on the event Ω_δ^m it holds that

$$\begin{aligned}\mathbb{E}_+(\|u-w\|^2) &\le e_n(u)^2 + 2(1-\delta)^{-2}\mathbb{E}_+(\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2) + 2(1-\delta)^{-2}\mathbb{E}_+(\|\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2)\\ &\le e_n(u)^2 + 2(1-\delta)^{-2}\mathbb{E}(\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2) + 2(1-\delta)^{-2}\mathbb{E}(\|\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2).\end{aligned}$$

Thanks to Lemma 3 in Section 5, we can bound the terms $\mathbb{E}(\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2)$ and $\mathbb{E}(\|\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2)$, and finally obtain the thesis. □

3.2 Convergence estimates in probability

A first bound in probability is given in the following theorem: it is valid with the general unbounded noise model (14).

Theorem 2. *For any $\alpha \in (0,1)$ and in the case of the noise model* (14) *with nonzero mean: if m satisfies the condition*

$$\frac{m}{\ln m + \ln(4\alpha^{-1})} \ge \frac{K(V_n)}{\zeta(\delta)}, \tag{33}$$

then it holds that

$$\begin{aligned}\Pr\Bigg(\|u-w\|^2 \le \Bigg(1 + \frac{4\zeta(\delta)}{\alpha(1-\delta)^2\ln(4m\alpha^{-1})}\Bigg) e_n(u)^2 \\ + \frac{4}{\alpha(1-\delta)^2}\left(\frac{\sigma^2 n}{m} + \|\overline{\eta}\|^2\left(1 + \frac{\zeta(\delta)}{\ln(4m\alpha^{-1})}\right)\right)\Bigg) > 1-\alpha.\end{aligned} \tag{34}$$

Proof. From Lemma 2 in Section 5 it holds that, for any $\phi > e_n(u)$

$$\begin{aligned}\Pr(\|u-w\| > \phi) &= \int_{\Omega_\delta^m} \mathbb{I}_{\{\|u-w\|>\phi\}} d\nu^m + \int_{\Omega^m \setminus \Omega_\delta^m} \mathbb{I}_{\{\|u-w\|>\phi\}} d\nu^m\\ &\le \int_{\Omega_\delta^m} \mathbb{I}_{\{e_n(u)^2 + 2(1-\delta)^{-2}\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2 + 2(1-\delta)^{-2}\|\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2 > \phi^2\}} d\nu^m + 2m^{-r}\\ &\le \int_{\Omega_\delta^m} \frac{2(1-\delta)^{-2}}{\phi^2 - e_n(u)^2}\left(\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2 + \|\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2\right) d\nu^m + 2m^{-r}\\ &\le \frac{2(1-\delta)^{-2}}{\phi^2 - e_n(u)^2}\left(\mathbb{E}_+(\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2) + \mathbb{E}_+(\|\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2)\right) + 2m^{-r}\\ &\le \frac{2(1-\delta)^{-2}}{\phi^2 - e_n(u)^2}\left(\mathbb{E}(\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2) + \mathbb{E}(\|\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2)\right) + 2m^{-r}\\ &\le \frac{2(1-\delta)^{-2}}{\phi^2 - e_n(u)^2}\left(\frac{\kappa(\delta)}{\ln m} e_n(u)^2 + \frac{n\sigma^2}{m} + \left(1 + \frac{\kappa(\delta)}{\ln m}\right)\|\overline{\eta}\|^2\right) + 2m^{-r},\end{aligned}$$

where we have used the results in Lemma 3 in Section 5 to bound the terms $\mathbb{E}(\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2)$ and $\mathbb{E}(\|\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2)$, and used the fact that condition (33) implies (19). Choosing now ϕ and r such that

$$\frac{2(1-\delta)^{-2}}{\phi^2 - e_n(u)^2}\left(\frac{\kappa(\delta)}{\ln m} e_n(u)^2 + \frac{n\sigma^2}{m} + \left(1 + \frac{\kappa(\delta)}{\ln m}\right)\|\overline{\eta}\|^2\right) = \frac{\alpha}{2}, \qquad 2m^{-r} = \frac{\alpha}{2},$$

leads to the final result. □

The probabilistic error bound (34) is not so satisfactory because of the factor $1/\alpha$ multiplying the different error terms. In the noiseless case $\eta = 0$, such factor can be easily eliminated by bounding the discrete least-squares error by the best approximation error in the L^∞ norm instead of the L^2_μ norm.

We now improve the result of Theorem 2 and in particular the factor $1/\alpha$ in front of the noise term by making more assumptions on the noise, namely, that $\widetilde{\eta}$ is bounded.

Theorem 3. *For any $\alpha \in (0,1)$ and in the case of the noise model (13) with nonzero mean: if m satisfies the condition*

$$\frac{m}{\ln m + \ln(6\alpha^{-1})} \geq \frac{K(V_n)}{\zeta(\delta)} \tag{35}$$

then

$$\begin{aligned}\Pr\Bigg(\|u-w\|^2 \leq \left(1+\frac{4\zeta(\delta)}{\alpha(1-\delta)^2\ln(6m\alpha^{-1})}\right)e_n(u)^2 \\ +\frac{8(1+\delta)}{(1-\delta)^2}\left(\widetilde{\eta}_{max}^2\ln\left(6m\alpha^{-1}\right)\frac{n}{m}\right)+\frac{8}{\alpha(1-\delta)^2}\left(1+\frac{\zeta(\delta)}{\ln(6m\alpha^{-1})}\right)\|\overline{\eta}\|^2\Bigg) > 1-\alpha.\end{aligned} \tag{36}$$

Proof. From (61), for any $f \in [0,1]$ and $\phi > e_n(u)$, it holds that

$$\begin{aligned}\Pr(\|u-w\|^2>\phi^2) &= \int_{\Omega_\delta^m}\mathbb{I}_{\{\|u-w\|^2>\phi^2\}}d\nu^m + \int_{\Omega^m\setminus\Omega_\delta^m}\mathbb{I}_{\{\|u-w\|^2>\phi^2\}}d\nu^m \\ &\leq \int_{\Omega_\delta^m}\mathbb{I}_{\left\{e_n(u)^2+2(1-\delta)^{-2}(\|\mathbf{J}^\top\mathbf{g}\|_{\ell^2}^2+2\|\mathbf{J}^\top\overline{\boldsymbol{\eta}}\|_{\ell^2}^2+2\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2)>\phi^2\right\}}d\nu^m + 2m^{-r} \\ &= \int_{\Omega_\delta^m}\mathbb{I}_{\left\{\|\mathbf{J}^\top\mathbf{g}\|_{\ell^2}^2+2\|\mathbf{J}^\top\overline{\boldsymbol{\eta}}\|_{\ell^2}^2+2\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2>\frac{\phi^2-e_n(u)^2}{2(1-\delta)^{-2}}\right\}}d\nu^m + 2m^{-r} \\ &\leq \int_{\Omega_\delta^m}\mathbb{I}_{\left\{\|\mathbf{J}^\top\mathbf{g}\|_{\ell^2}^2+2\|\mathbf{J}^\top\overline{\boldsymbol{\eta}}\|_{\ell^2}^2>f\frac{\phi^2-e_n(u)^2}{2(1-\delta)^{-2}}\right\}}d\nu^m + \int_{\Omega_\delta^m}\mathbb{I}_{\left\{\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2>(1-f)\frac{\phi^2-e_n(u)^2}{4(1-\delta)^{-2}}\right\}}d\nu^m + 2m^{-r} \\ &\leq \int_{\Omega_\delta^m}\frac{2(1-\delta)^{-2}\left(\|\mathbf{J}^\top\mathbf{g}\|_{\ell^2}^2+2\|\mathbf{J}^\top\overline{\boldsymbol{\eta}}\|_{\ell^2}^2\right)}{f(\phi^2-e_n(u)^2)}d\nu^m + \int_{\Omega_\delta^m}\mathbb{I}_{\left\{\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2>(1-f)\frac{\phi^2-e_n(u)^2}{4(1-\delta)^{-2}}\right\}}d\nu^m + 2m^{-r} \\ &= \int_{\Omega_\delta^m}\frac{2(1-\delta)^{-2}\left(\|\mathbf{J}^\top\mathbf{g}\|_{\ell^2}^2+2\|\mathbf{J}^\top\overline{\boldsymbol{\eta}}\|_{\ell^2}^2\right)}{f(\phi^2-e_n(u)^2)}d\nu^m + \Pr(\Omega_\delta^m)\Pr\left(\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2>(1-f)\frac{\phi^2-e_n(u)^2}{4(1-\delta)^{-2}}\middle|\Omega_\delta^m\right)+2m^{-r} \\ &\leq \frac{2(1-\delta)^{-2}\left(\mathbb{E}\left(\|\mathbf{J}^\top\mathbf{g}\|_{\ell^2}^2\right)+2\mathbb{E}\left(\|\mathbf{J}^\top\overline{\boldsymbol{\eta}}\|_{\ell^2}^2\right)\right)}{f(\phi^2-e_n(u)^2)} + \Pr\left(\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2>(1-f)\frac{\phi^2-e_n(u)^2}{4(1-\delta)^{-2}}\middle|\Omega_\delta^m\right)+2m^{-r} \\ &\leq \frac{(1-\delta)^{-2}(\epsilon(m,\delta,r)e_n(u)^2+2\left(4+\epsilon(m,\delta,r)\right)\|\overline{\eta}\|^2)}{2f(\phi^2-e_n(u)^2)} \\ &\quad + \Pr\left(\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2>(1-f)\frac{\phi^2-e_n(u)^2}{4(1-\delta)^{-2}}\middle|\Omega_\delta^m\right)+2m^{-r},\end{aligned}$$

where we have used the results in Lemma 3 in Section 5 to bound the terms $\mathbb{E}(\|\mathbf{J}^\top\mathbf{g}\|_{\ell^2}^2)$ and $\mathbb{E}(\|\mathbf{J}^\top\overline{\boldsymbol{\eta}}\|_{\ell^2}^2)$, and again the fact that condition (35) implies (19). Now we choose

$$\phi^2 = \frac{8(1-\delta)^{-2}}{1-f}(1+r)(1+\delta)\widetilde{\eta}_{\max}^2 n\frac{\ln m}{m} + e_n(u)^2, \tag{37}$$

such that, by applying Theorem 9, it holds that

$$\Pr\left(\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2>(1-f)\frac{\phi^2-e_n(u)^2}{4(1-\delta)^{-2}}\middle|\Omega_\delta^m\right)\leq m^{-r}.$$

We choose

$$r = -\frac{\ln(\alpha/6)}{\ln m} \tag{38}$$

such that

$$3m^{-r} = \frac{\alpha}{2},$$

and f such that

$$\frac{(1-\delta)^{-2}\left(\epsilon(m,\delta,r)e_n(u)^2+2\left(4+\epsilon(m,\delta,r)\right)\|\overline{\eta}\|^2\right)}{2f(\phi^2-e_n(u)^2)} = \frac{\alpha}{2}. \tag{39}$$

By replacing the expression of ϕ from (37) into (39) we obtain that

$$f = \frac{m\left(\epsilon(m,\delta,r)e_n(u)^2 + 2\left(4+\epsilon(m,\delta,r)\right)\|\overline{\eta}\|^2\right)}{8\alpha(1+r)(1+\delta)\widetilde{\eta}_{\max}^2 n\ln m + m\left(\epsilon(m,\delta,r)e_n(u)^2 + 2\left(4+\epsilon(m,\delta,r)\right)\|\overline{\eta}\|^2\right)}. \tag{40}$$

Notice that it is always true that $0 < f < 1$. The thesis follows by choosing ϕ, r and f as in (37), (38) and (40), respectively. □

In the next theorem we completely eliminate the factor $1/\alpha$, but at the price of introducing the L^∞ best approximation error in the estimate.

Theorem 4. *For any $\alpha \in (0,1)$ and with the noise model (12) with nonzero mean: if m satisfies the condition*

$$\frac{m}{\ln m + \ln(4\alpha^{-1})} \geq \frac{K(V_n)}{\zeta(\delta)}$$

then

$$\begin{aligned}\Pr\Bigg(\|u-w\|^2 \leq e_n(u)^2 + \frac{2}{1-\delta}e_n^\infty(u)^2 \\ + \frac{8(1+\delta)}{(1-\delta)^{-2}}\left(\widetilde{\eta}_{max}^2 \ln\left(4m\alpha^{-1}\right)\frac{n}{m}\right) + \frac{4(1+\delta)}{(1-\delta)^2}\overline{\eta}_{max}^2\Bigg) > 1-\alpha.\end{aligned} \tag{41}$$

Proof. From (62), for any $f \in [0,1]$, it holds that

$$\begin{aligned}\Pr(\|u-w\|^2 > \phi^2) &= \int_{\Omega_\delta^m} \mathbb{I}_{\{\|u-w\|^2>\phi^2\}} d\nu^m + \int_{\Omega^m\setminus\Omega_\delta^m} \mathbb{I}_{\{\|u-w\|^2>\phi^2\}} d\nu^m \\ &\leq \int_{\Omega_\delta^m} \mathbb{I}_{\left\{e_n(u)^2+4(1-\delta)^{-2}\left(\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2+(1+\delta)\overline{\eta}_{\max}^2\right)+2(1-\delta)^{-1}e_n^\infty(u)^2>\phi^2\right\}} d\nu^m + 2m^{-r} \\ &= \int_{\Omega_\delta^m} \mathbb{I}_{\left\{\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2 > \frac{\phi^2 - e_n(u)^2 - 2(1-\delta)^{-1}\left(e_n^\infty(u)^2+2(1-\delta)^{-1}(1+\delta)\overline{\eta}_{\max}^2\right)}{4(1-\delta)^{-2}}\right\}} d\nu^m + 2m^{-r} \\ &= \Pr(\Omega_\delta^m)\Pr\left(\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2 > \frac{\phi^2 - e_n(u)^2 - 2(1-\delta)^{-1}\left(e_n^\infty(u)^2+2(1-\delta)^{-1}(1+\delta)\overline{\eta}_{\max}^2\right)}{4(1-\delta)^{-2}}\middle|\Omega_\delta^m\right) + 2m^{-r}.\end{aligned}$$

Now we choose

$$\phi^2 = 8(1+r)(1-\delta)^{-2}(1+\delta)\widetilde{\eta}_{\max}^2 n\frac{\ln m}{m} + e_n(u)^2 + 2(1-\delta)^{-1}\left(e_n^\infty(u)^2 + 2(1-\delta)^{-1}(1+\delta)\overline{\eta}_{\max}^2\right), \tag{42}$$

such that, by applying Theorem 9, it holds that

$$\Pr\left(\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2 > \frac{\phi^2 - e_n(u)^2 - 2(1-\delta)^{-1}\left(e_n^\infty(u)^2+2(1-\delta)^{-1}(1+\delta)\overline{\eta}_{\max}^2\right)}{4(1-\delta)^{-2}}\middle|\Omega_\delta^m\right) \leq 2m^{-r}.$$

We choose

$$r = -\frac{\ln(\alpha/4)}{\ln m} \tag{43}$$

such that

$$4m^{-r} = \alpha,$$

The thesis follows by choosing ϕ and r as in (42) and (43), respectively. □

A corollary of Theorem 4 is the following, where only the L^∞ best approximation error is used.

Corollary 1. *For any $\alpha \in (0,1)$ and with the noise model (12) with nonzero mean: if the condition*

$$\frac{m}{\ln m + \ln 2\alpha^{-1}} \geq \frac{K(V_n)}{\zeta(\delta)}$$

holds true then

$$\Pr\left(\|u - \Pi_n^m u\|^2 \leq (1 + 2(1-\delta)^{-1}) e_n^\infty(u)^2 + 4(1-\delta)^{-2}(1+\delta)\eta_{max}^2\right) > 1 - \alpha. \tag{44}$$

Proof. Following the lines of the proof of Theorem 4, we choose $\widetilde{\eta} = 0$ and

$$r = -\frac{\ln(\alpha/2)}{\ln m} \tag{45}$$

since

$$2m^{-r} = \alpha.$$

Of course it holds $e_n(u) \leq e_n^\infty(u)$, and we obtain the thesis. □

In the noisy case, a similar result to Corollary 1 is stated in Theorem 2.2 in [4] for the deterministic noise model.

4 The case of polynomial approximation $V_n = \mathbb{P}_\Lambda$

In the following, we present further results concerning our analysis of the stability and accuracy properties of the discrete L^2 projection, in the specific case of multivariate polynomial approximation spaces.

We now introduce further information concerning the structure of the marginal probability μ. Given a collection of (possibly different) univariate measures $\mu_i : \Gamma_i \to \mathbb{R}_0^+$ for any $i = 1, \ldots, d$, we assume that μ can be expressed as a product measure $\mu(y) = \prod_{i=1}^d \mu(y_i)$. For any $i = 1, \ldots, d$, consider the (possibly different) measures μ_i defined on the interval Γ_i, with corresponding densities $\rho_i : \Gamma_i \to \mathbb{R}^+$. We introduce the family $(\varphi_k^i)_{k\geq 0}$ of $L^2_{\mu_i}$-orthonormal polynomials of degree k, *i.e.* these polynomials are orthonormal w.r.t. the weighted L^2 inner product (4) with the weight being the probability density function ρ_i associated with the measure μ_i:

$$\int_{\Gamma_i} \varphi_j^i(y)\varphi_k^i(y) d\mu_i(y) = \int_{\Gamma_i} \varphi_j^i(y)\varphi_k^i(y)\rho_i(y) d\lambda(y) = \int_{\Gamma_i} \varphi_j^i(y)\varphi_k^i(y)\rho_i(y) dy = \delta_{jk}.$$

We introduce the gamma function $\Gamma(\theta) := \int_0^{+\infty} t^{\theta-1}e^{-t}dt$ with $\mathrm{Re}(\theta) > 0$ then extended by analytic continuation, and the beta function $\mathcal{B}(\theta_1, \theta_2) := \Gamma(\theta_1)\Gamma(\theta_2)/\Gamma(\theta_1+\theta_2)$ for any $\theta_1, \theta_2 > -1$. We focus on the univariate Jacobi weight,

$$\rho_J^{\theta_1,\theta_2}(y) = \left(2^{\theta_1+\theta_2+1}\mathcal{B}(\theta_1+1, \theta_2+1)\right)^{-1}(1-y)^{\theta_1}(1+y)^{\theta_2}, \quad y \in [-1,1], \tag{46}$$

with real shape parameters $\theta_1, \theta_2 > -1$ which leads to the family of univariate Jacobi polynomials $(J_k^{\theta_1,\theta_2})_{k\geq 0}$. Remarkable instances of Jacobi polynomials are Legendre polynomials when $\theta_1 = \theta_2 = 0$, and Chebyshev polynomials of the first kind when $\theta_1 = \theta_2 = -1/2$. Notice that the weight (46) is normalized such that it integrates to one over the whole support, also known as probabilistic orthonormalization. The Jacobi weight $\rho_J^{\theta_1,\theta_2}$ corresponds, up to a translation in the parameters θ_1, θ_2 and in the support, to the standard beta probability density function.

Given a d-dimensional set $\Lambda \subseteq \mathcal{F} := \mathbb{N}_0^d$ of multi-indices, we define the polynomial space $\mathbb{P}_\Lambda = \mathbb{P}_\Lambda(\Gamma)$ as

$$\mathbb{P}_\Lambda := \mathrm{span}\{\psi_{\boldsymbol{q}} : \boldsymbol{q} \in \Lambda\},$$

with each multivariate polynomial basis function being defined as

$$\psi_{\boldsymbol{q}}(y) := \prod_{i=1}^{d} \varphi_{q_i}^i(y_i), \quad y \in \Gamma, \tag{47}$$

by tensorization of the univariate families of $L^2_{\rho_i}$-orthonormal polynomials. This setting can be extended to the case $d = +\infty$ by taking $\Lambda = \mathcal{F}$, *i.e.* the set of finitely supported sequences $(q_1, q_2, \ldots)$ and observing that $\varphi_0 = 1$ for any probability density function ρ, so that the product in (47) has only a finite number of factors different than 1. We denote the cardinality of the multi-index set Λ by $\#\Lambda$.

The discrete seminorm becomes a norm almost surely over $\mathbb{P}_\Lambda$, provided the points are distinct and their number satisfies $m \geq \dim(\mathbb{P}_\Lambda)$.

In the case of polynomial approximation we set $V_n = \mathbb{P}_\Lambda$ with $n = \dim(\mathbb{P}_\Lambda) = \#\Lambda$.

Definition 1 (Ordering $\leq$ for multi-indices)**.** *For any $\boldsymbol{q}, \boldsymbol{p} \in \mathcal{F}$, the ordering $\boldsymbol{q} \leq \boldsymbol{p}$ means that $q_i \leq p_i$ for any $i = 1, \ldots, d$.*

Definition 2 (Downward closed multi-index set)**.** *In any dimension d, a multi-index set $\Lambda \subset \mathcal{F}$ is downward closed (or it is a lower set) if*

$$\boldsymbol{q} \in \Lambda \implies \boldsymbol{p} \in \Lambda, \ \forall \boldsymbol{p} \leq \boldsymbol{q}.$$

The following lemma merges the results proven in [4, 13] concerning upper bounds on the quantity K defined in (18) for Jacobi polynomials. The polynomial space $\mathbb{P}_\Lambda$ is uniquely determined from the multi-index set Λ, and therefore we shorten the notation to $K(\Lambda) = K(\mathbb{P}_\Lambda) = K(V_n)$.

Lemma 1. *In any dimension d, for any downward closed set $\Lambda \subset \mathcal{F}$ the quantity K with Jacobi polynomials $(J_k^{\theta_1,\theta_2})_{k\geq 0}$ and any $\theta_1, \theta_2 \in \mathbb{N}_0$ satisfies*

$$K(\Lambda) \leq (\#\Lambda)^{2\max\{\theta_1,\theta_2\}+2}, \tag{48}$$

and with Chebyshev polynomials of the first kind $(J_k^{\theta_1,\theta_2})_{k\geq 0}$ and $\theta_1 = \theta_2 = -1/2$ satisfies

$$K(\Lambda) \leq (\#\Lambda)^{\ln 3/\ln 2}. \tag{49}$$

In particular for Legendre polynomials $(J_k^{\theta_1,\theta_2})_{k\geq 0}$ and $\theta_1 = \theta_2 = 0$ it holds

$$K(\Lambda) \leq (\#\Lambda)^2.$$

We denote by Π_Λ^m and Π_Λ the discrete and continuous L^2 projections in the case of polynomial approximation. In what follows, we state the convergence results for the discrete least-squares approximation in expectation, both in the noiseless case (from [4]) and in the noisy case as a consequence of Theorem 1, and the results in probability, which are consequences of Theorems 2, 3, 4, Corollary 1 and [4, Theorem 3] in the noiseless case.

Theorem 5. *For any $r > 0$, any $\delta \in (0, 1)$ and any downward closed multi-index set $\Lambda \subset \mathbb{N}_0^d$, if m satisfies*

$$\frac{m}{\ln m} \geq \frac{1+r}{\zeta(\delta)} (\#\Lambda)^{2\max\{\theta_1,\theta_2\}+2} \tag{50}$$

with Jacobi polynomials $(J_k^{\theta_1,\theta_2})_{k\geq 0}$ and any $\theta_1, \theta_2 \in \mathbb{N}_0$, or if m satisfies

$$\frac{m}{\ln m} \geq \frac{1+r}{\zeta(\delta)} (\#\Lambda)^{(\ln 3/\ln 2)} \tag{51}$$

with Chebyshev polynomials of the first kind $(J_k^{\theta_1,\theta_2})_{k\geq 0}$ *and* $\theta_1 = \theta_2 = -1/2$*, then for any* $u \in L^\infty(\Gamma)$ *with* $\|u\|_{L^\infty(\Gamma)} \leq \tau$*, the following holds: in the case of the noiseless model*

$$\mathbb{E}\left(\|u - \widetilde{\Pi}_\Lambda^m u\|^2\right) \leq \left(1 + \frac{4\zeta(\delta)}{(1+r)\ln m}\right)\|u - \Pi_\Lambda u\|^2 + 8\tau^2 m^{-r}, \tag{52}$$

and in the case of the noise model (14) *with nonzero mean*

$$\begin{aligned}\mathbb{E}(\|u - \widetilde{w}\|^2) \leq &\left(1 + \frac{2\zeta(\delta)}{(1-\delta)^2(1+r)\ln m}\right) e_n(u)^2 \\ &+ \frac{2}{(1-\delta)^2}\left(\frac{\sigma^2 \#\Lambda}{m} + \|\overline{\eta}\|^2\left(1 + \frac{\zeta(\delta)}{(1+r)\ln m}\right)\right) + 8\tau^2 m^{-r}.\end{aligned} \tag{53}$$

Theorem 6. *For any* $\alpha \in (0,1)$*, any* $s \geq 1$*, any* $\delta \in (0,1)$ *and any downward closed multi-index set* $\Lambda \subset \mathbb{N}_0^d$*, consider the following conditions between* m *and* $\#\Lambda$*:*

$$\frac{m}{\ln m + \ln(s\alpha^{-1})} \geq \frac{1}{\zeta(\delta)}(\#\Lambda)^{2\max\{\theta_1,\theta_2\}+2} \tag{54}$$

with Jacobi polynomials $(J_k^{\theta_1,\theta_2})_{k\geq 0}$ *and any* $\theta_1, \theta_2 \in \mathbb{N}_0$*, or*

$$\frac{m}{\ln m + \ln(s\alpha^{-1})} \geq \frac{1}{\zeta(\delta)}(\#\Lambda)^{(\ln 3/\ln 2)} \tag{55}$$

with Chebyshev polynomials of the first kind $(J_k^{\theta_1,\theta_2})_{k\geq 0}$ *and* $\theta_1 = \theta_2 = -1/2$*.*

- *In the case of the noiseless model, under condition* (54) *or* (55) *with* $s = 2$ *it holds that*

$$\begin{aligned}&\Pr\left(\|u - \Pi_\Lambda^m u\| \leq \left(1 + \sqrt{\frac{1}{1-\delta}}\right) e_n^\infty(u)\right) \geq 1 - \alpha, \\ &\Pr\left(cond(\mathbf{G}) \leq \frac{1+\delta}{1-\delta}\right) \geq 1 - \alpha.\end{aligned}$$

- *In the case of the noise model* (14) *with nonzero mean, under condition* (54) *or* (55) *with* $s = 4$ *it holds that*

$$\begin{aligned}\Pr\Bigg(\|u - w\|^2 \leq &\left(1 + \frac{4\zeta(\delta)}{\alpha(1-\delta)^2\ln(4m\alpha^{-1})}\right) e_n(u)^2 \\ &+ \frac{4}{\alpha(1-\delta)^2}\left(\frac{\sigma^2\#\Lambda}{m} + \|\overline{\eta}\|^2\left(1 + \frac{\zeta(\delta)}{\ln(4m\alpha^{-1})}\right)\right)\Bigg) > 1 - \alpha.\end{aligned} \tag{56}$$

- *In the case of the noise model* (13) *with nonzero mean, under condition* (54) *or* (55) *with* $s = 6$ *it holds that,*

$$\begin{aligned}\Pr\Bigg(\|u - w\|^2 \leq &\left(1 + \frac{4\zeta(\delta)}{\alpha(1-\delta)^2\ln(6m\alpha^{-1})}\right) e_n(u)^2 + \frac{8(1+\delta)}{(1-\delta)^2}\left(\widetilde{\eta}_{max}^2 \ln\left(6m\alpha^{-1}\right)\frac{\#\Lambda}{m}\right) \\ &+ \frac{8}{\alpha(1-\delta)^2}\left(1 + \frac{\zeta(\delta)}{\ln(6m\alpha^{-1})}\right)\|\overline{\eta}\|^2\Bigg) > 1 - \alpha.\end{aligned} \tag{57}$$

- *In the case of the noise model* (12) *with nonzero mean, under condition* (54) *or* (55) *with* $s = 4$ *it holds that*

$$\begin{aligned}\Pr\Bigg(\|u - w\|^2 \leq &\, e_n(u)^2 + \frac{2}{1-\delta}e_n^\infty(u)^2 \\ &+ \frac{8(1+\delta)}{(1-\delta)^{-2}}\left(\widetilde{\eta}_{max}^2 \ln\left(4m\alpha^{-1}\right)\frac{\#\Lambda}{m}\right) + \frac{4(1+\delta)}{(1-\delta)^2}\overline{\eta}_{max}^2\Bigg) > 1 - \alpha.\end{aligned} \tag{58}$$

- *In the case of the noise model* (12) *with nonzero mean, under condition* (54) *or* (55) *with* $s = 2$ *it holds that*

$$\Pr\left(\|u - \Pi_n^m u\|^2 \leq (1 + 2(1-\delta)^{-1})e_n^\infty(u) + 4(1-\delta)^{-2}(1+\delta)\eta_{max}^2\right) > 1 - \alpha. \tag{59}$$

5 Intermediate results

This section collects some intermediate results that have been used to prove our convergence estimates in Section 3.

Lemma 2. *On the event* Ω_δ^m *it holds that*

$$\begin{aligned}\|u - w\|^2 \leq & e_n(u)^2 + 2(1-\delta)^{-2}\left(\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2 + \|\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2\right) &(60)\\ \leq & e_n(u)^2 + 2(1-\delta)^{-2}\left(\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2 + 2\|\mathbf{J}^\top \overline{\boldsymbol{\eta}}\|_{\ell^2}^2 + 2\|\mathbf{J}^\top \widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2\right) &(61)\end{aligned}$$

and, in the case of bounded offset,

$$\|u - w\|^2 \leq e_n(u)^2 + 4(1-\delta)^{-2}\left(\|\mathbf{J}^\top \widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2 + (1+\delta)\overline{\eta}_{max}^2\right) + 2(1-\delta)^{-1}e_n^\infty(u)^2. \tag{62}$$

Proof. On the event Ω_δ^m we have

$$\begin{aligned}\|u - w\|^2 &= \|u - \Pi_n u - \Pi_n^m(u - \Pi_n u) + \Pi_n^m u - w\|^2\\ &\leq \|u - \Pi_n u\|^2 + 2\|\Pi_n^m(u - \Pi_n u)\|^2 + 2\|\Pi_n^m u - w\|^2 &(63)\\ &= e_n(u)^2 + 2\|\mathbf{G}^{-1}\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2 + 2\|\mathbf{G}^{-1}\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2\\ &\leq e_n(u)^2 + 2(1-\delta)^{-2}\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2 + 2(1-\delta)^{-2}\|\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2 &(64)\\ &\leq e_n(u)^2 + 2(1-\delta)^{-2}\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2 + 4(1-\delta)^{-2}\left(\|\mathbf{J}^\top \overline{\boldsymbol{\eta}}\|_{\ell^2}^2 + \|\mathbf{J}^\top \widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2\right),\end{aligned}$$

where $\mathbf{g}$ is the vector with elements defined in (6). At (64) we have proven (60). Then we have exploited the upper bound in (26), the splitting (9), the triangular inequality and $(a+b)^2 \leq 2(a^2+b^2)$, $\forall a, b \geq 0$ to obtain (61). To prove (62), we start from (63), estimate the term

$$\|\Pi_n^m(u - \Pi_n u)\|^2 \leq (1-\delta)^{-1}\|\Pi_n^m u - \Pi_n u\|_m^2 \leq (1-\delta)^{-1}\|u - \Pi_n u\|_m^2 \leq (1-\delta)^{-1}\|u - \Pi_n u\|_{L^\infty}^2$$

using (23) and estimate the term

$$\begin{aligned}\|\Pi_n^m u - w\|^2 &\leq 2\|\Pi_n^m u - w + \Pi_n^m \overline{\eta}\|^2 + 2\|\Pi_n^m \overline{\eta}\|^2\\ &= 2\|\mathbf{G}^{-1}\mathbf{J}^\top \widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2 + 2\|\mathbf{G}^{-1}\mathbf{J}^\top \overline{\boldsymbol{\eta}}\|_{\ell^2}^2\\ &\leq 2(1-\delta)^{-2}\|\mathbf{J}^\top \widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2 + 2(1-\delta)^{-2}\|\mathbf{J}^\top \overline{\boldsymbol{\eta}}\|_{\ell^2}^2\\ &\leq 2(1-\delta)^{-2}\|\mathbf{J}^\top \widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2 + 2(1-\delta)^{-2}|||\mathbf{J}^\top|||^2\|\overline{\boldsymbol{\eta}}\|^2\\ &\leq 2(1-\delta)^{-2}\|\mathbf{J}^\top \widetilde{\boldsymbol{\eta}}\|_{\ell^2}^2 + 2(1-\delta)^{-2}(1+\delta)\overline{\eta}_{\max}^2\end{aligned}$$

using the splitting (9), (26) and (27). □

Lemma 3. *Under condition* (19) *it holds that*

$$\mathbb{E}(\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2) \leq \frac{\kappa(\delta)}{\ln m} e_n(u)^2, \tag{65}$$

$$\mathbb{E}(\|\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2) \leq \frac{n\sigma^2}{m} + \left(1 + \frac{\kappa(\delta)}{\ln m}\right)\|\overline{\boldsymbol{\eta}}\|^2, \tag{66}$$

$$\mathbb{E}(\|\mathbf{J}^\top \overline{\boldsymbol{\eta}}\|_{\ell^2}^2) \leq \left(1 + \frac{\kappa(\delta)}{\ln m}\right)\|\overline{\eta}\|^2. \tag{67}$$

Proof. The term $\mathbb{E}(\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2)$ is estimated in [5, Theorem 2] and it holds

$$\mathbb{E}(\|\mathbf{J}^\top \mathbf{g}\|_{\ell^2}^2|) \leq \frac{K(V_n)}{m}\|u - \Pi_n u\|^2 \leq \frac{\kappa(\delta)}{\ln m} e_n(u)^2.$$

For the term $\mathbb{E}(\|\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2)$, following [5, Theorem 3] and denoting $\widetilde{\eta}_j := \eta_j - \overline{\eta}(y_j)$ for any $j = 1, \ldots, m$, we have

$$\begin{aligned}
\mathbb{E}(\|\mathbf{J}^\top \boldsymbol{\eta}\|_{\ell^2}^2) &= \mathbb{E}\left(\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m}\eta_j\psi_i(y_j)\right)^2\right) \\
&= \sum_{i=1}^{n}\frac{1}{m^2}\left(\sum_{j=1}^{m}\mathbb{E}((\eta_j)^2\psi_i(y_j)^2) + \sum_{j\neq k}\mathbb{E}\left(\eta_j\eta_k\psi_i(y_j)\psi_i(y_k)\right)\right) \\
&= \sum_{i=1}^{n}\frac{1}{m^2}\left(\sum_{j=1}^{m}\mathbb{E}((\widetilde{\eta}_j)^2\psi_i(y_j)^2) + \sum_{j=1}^{m}\mathbb{E}(\overline{\eta}(y_j)^2\psi_i(y_j)^2) + \sum_{j\neq k}\mathbb{E}\left(\overline{\eta}(y_j)\overline{\eta}(y_k)\psi_i(y_j)\psi_i(y_k)\right)\right) \\
&= \sum_{i=1}^{n}\frac{1}{m^2}\left(m\mathbb{E}(\mathbb{E}(\widetilde{\eta}^2|y)\psi_i(y)^2) + m\mathbb{E}(\overline{\eta}^2\psi_i^2) + m(m-1)\mathbb{E}(\overline{\eta}\psi_i)^2\right) \\
&\leq \frac{n\sigma^2}{m} + \left(\frac{K(V_n)}{m} + 1 - \frac{1}{m}\right)\|\overline{\eta}\|^2 \\
&\leq \frac{n\sigma^2}{m} + \left(1 + \frac{\kappa(\delta)}{\ln m}\right)\|\overline{\eta}\|^2.
\end{aligned}$$

As a minor product from above we obtain

$$\begin{aligned}
\mathbb{E}(\|\mathbf{J}^\top \overline{\boldsymbol{\eta}}\|_{\ell^2}^2) &= \mathbb{E}\left(\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m}\overline{\eta}_j\psi_i(y_j)\right)^2\right) \\
&= \sum_{i=1}^{n}\frac{1}{m^2}\left(\sum_{j=1}^{m}\mathbb{E}((\overline{\eta}_j)^2\psi_i(y_j)^2) + \sum_{j\neq k}\mathbb{E}(\overline{\eta}_j\overline{\eta}_k\psi_i(y_j)\psi_i(y_k)\right) \\
&= \sum_{i=1}^{n}\frac{1}{m^2}\left(m\mathbb{E}(\overline{\eta}^2\psi_i^2) + m(m-1)\mathbb{E}(\overline{\eta}\psi_i)^2\right) \\
&\leq \left(\frac{K(V_n)}{m} + 1 - \frac{1}{m}\right)\|\overline{\eta}\|^2 \\
&\leq \left(1 + \frac{\kappa(\delta)}{\ln m}\right)\|\overline{\eta}\|^2.
\end{aligned}$$

□

5.1 A large deviation estimate for the noise term $\mathbf{J}^\top\boldsymbol{\eta}$

In this section we analyze the noise term $\mathbf{J}^\top\boldsymbol{\eta}$ with arguments coming from the theory of large deviations [8, 21]. The main result of this section is Theorem 9. The final goal is achieved in Theorem 3 and Theorem 4, which contain a result in probability for the noisy case similar to Theorem 2, but it is proven using Theorem 9 instead of (66) in Lemma 3 to bound the term $\mathbb{E}(\|\mathbf{J}^\top\boldsymbol{\eta}\|_{\ell^2}^2)$. We start with a technical lemma.

Lemma 4. *For any $L \geq 0$, given n random variables $\mathcal{M}_1, \ldots, \mathcal{M}_n$ and for any nonnegative reals $l_1, \ldots, l_n$ such that $\sum_{k=1}^n l_k = L$ it holds that*

$$\Pr\left(\sum_{k=1}^n \mathcal{M}_k^2 > L\right) \leq \sum_{k=1}^n \Pr\left(\mathcal{M}_k^2 > l_k\right). \tag{68}$$

Proof. Indeed we have the following inclusions between probability events:

$$\left\{\sum_{k=1}^n \mathcal{M}_k^2 < \sum_{k=1}^n l_k\right\} \supseteq \left\{\mathcal{M}_k^2 < l_k,\ \forall\, k \in \{1, \ldots, n\}\right\} = \bigcap_{k=1}^n \left\{\mathcal{M}_k^2 < l_k\right\},$$

and we conclude by switching to complementary events and using the subadditivity property of probability. □

Notice that, in Lemma 4, the random variables are not assumed to be independent. In the proof of Theorem 9 we use Lemma 4 with random variables that are mutually dependent on each other. We recall now the standard Hoeffding's inequality and a conditional version of it that will be used in the proof of Theorem 9.

Theorem 7 (Hoeffding's inequality)**.** *For any $t \geq 0$, given m independent random variables $\mathcal{X}_1, \ldots, \mathcal{X}_m$ almost surely bounded,* i.e. $\Pr(\mathcal{X}_j \in [a_j, b_j]) = 1$ *with $b_j \geq a_j$ for any $j = 1, \ldots, m$, their empirical mean*

$$\overline{\mathcal{X}} = \frac{1}{m}\sum_{j=1}^m \mathcal{X}_j$$

satisfies

$$\Pr\left(|\overline{\mathcal{X}} - \mathbb{E}[\overline{\mathcal{X}}]| \geq t\right) \leq 2\exp\left\{-\frac{2m^2t^2}{\sum_{j=1}^m (b_j - a_j)^2}\right\}. \tag{69}$$

Theorem 8 (Conditional Hoeffding's inequality)**.** *Let $(\Omega, \Sigma, \Pr)$ be a probability space and $\mathcal{X}_1, \ldots, \mathcal{X}_m, \mathbf{y}$ random variables measurable in $(\Omega, \Sigma, \Pr)$. We assume that $\mathcal{X}_j|\mathbf{y}$ are almost surely bounded,* i.e. $\Pr(\mathcal{X}_j|\mathbf{y} \in [a_j(\mathbf{y}), b_j(\mathbf{y})]) = 1$ *for some $b_j(\mathbf{y}) \geq a_j(\mathbf{y})$, and are conditionally independent. Then, for any $t \geq 0$ the (conditional) empirical mean*

$$\overline{\mathcal{X}} = \frac{1}{m}\sum_{j=1}^m \mathcal{X}_j$$

satisfies

$$\Pr\left(|\overline{\mathcal{X}} - \mathbb{E}[\overline{\mathcal{X}}]| \geq t|\mathbf{y}\right) \leq 2\exp\left\{-\frac{2m^2t^2}{\sum_{j=1}^m (b_j(\mathbf{y}) - a_j(\mathbf{y}))^2}\right\}. \tag{70}$$

Now we present a result that allows us to bound in probability the zeromean noise term.

Theorem 9. *For any $r > 0$, any $\delta \in (0, 1)$ and any m satisfying condition* (19)*, for zeromean noise random variables $\widetilde{\boldsymbol{\eta}}$ satisfying the assumptions of the bounded noise model* (12)*, it holds that*

$$\Pr\left(\|\mathbf{J}^\top \widetilde{\boldsymbol{\eta}}\|_2^2 > 2(1+r)(1+\delta)\frac{\widetilde{\eta}_{max}^2 n \ln m}{m}\middle|\Omega_\delta^m\right) \leq 2m^{-r}. \tag{71}$$

Proof. The zeromean noise term can be written as

$$\|\mathbf{J}^\top \widetilde{\boldsymbol{\eta}}\|_2^2 = \sum_{k=1}^n \mathcal{M}_k^2,$$

with

$$\mathcal{M}_k := \frac{1}{m}\sum_{j=1}^{m} \psi_k(y_j)\widetilde{\eta}_j = (\psi_k, \widetilde{\boldsymbol{\eta}})_m.$$

For any $k = 1, \ldots, n$, each one of the random variables $\mathcal{M}_k|\mathbf{y}$ gives the empirical mean of the random variables $\mathcal{X}_{k,j}|\mathbf{y} := \psi_k(y_j)\widetilde{\eta}_j$ for $j = 1, \ldots, m$. Notice that all these random variables are conditioned to the points $\mathbf{y}$. Under the assumption of bounded noise (12), the random variables in the vector $\widetilde{\boldsymbol{\eta}}$ are bounded almost surely. Thus the random variables $\mathcal{X}_{k,j}|\mathbf{y}$, satisfy the bounds

$$\mathcal{X}_{k,j}|\mathbf{y} \in [a_{kj}(\mathbf{y}), b_{kj}(\mathbf{y})] := [-|\psi_k(y_j)|\widetilde{\eta}_{\max}, |\psi_k(y_j)|\widetilde{\eta}_{\max}], \quad \forall k = 1, \ldots, n, \forall j = 1, \ldots, m.$$

Since by construction $\mathbb{E}[\widetilde{\boldsymbol{\eta}}|\mathbf{y}] = 0$ then $\mathbb{E}[\mathcal{M}_k|\mathbf{y}] = 0$ for any $k = 1, \ldots, n$. Using the conditional Hoeffding's inequality (70) for each one of the random variables $\mathcal{M}_k|\mathbf{y}$ and the definition of discrete norm we obtain

$$\begin{aligned}
\Pr\left(|\mathcal{M}_k|^2 > l_k \middle| \mathbf{y}\right) =& \Pr\left(|\mathcal{M}_k| > \sqrt{l_k} \middle| \mathbf{y}\right) \\
\leq& 2\exp\left\{-\frac{2l_k m^2}{\sum_{j=1}^{m}(b_{kj}(\mathbf{y}) - a_{kj}(\mathbf{y}))^2}\right\} \\
=& 2\exp\left\{-\frac{2l_k m^2}{4\widetilde{\eta}_{\max}^2 \sum_{j=1}^{m}|\psi_k(y_j)|^2}\right\} \\
=& 2\exp\left\{-\frac{l_k m}{2\widetilde{\eta}_{\max}^2 \|\psi_k\|_m^2}\right\}. \qquad (72)
\end{aligned}$$

Then we use in sequence (16), (68), (72) and (28):

$$\begin{aligned}
\Pr\left(\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_2^2 > L \middle| \Omega_\delta^m\right) =& \Pr\left(\sum_{k=1}^{n}\mathcal{M}_k^2 > L \middle| \Omega_\delta^m\right) \\
=& \Pr\left(\sum_{k=1}^{n}\mathcal{M}_k^2 > \frac{L}{\mathrm{tr}(\mathbf{G})}\sum_{k=1}^{n}\|\psi_k\|_m^2 \middle| \Omega_\delta^m\right) \\
\leq& \sum_{k=1}^{n}\Pr\left(\mathcal{M}_k^2 > \frac{L}{\mathrm{tr}(\mathbf{G})}\|\psi_k\|_m^2 \middle| \Omega_\delta^m\right) \\
=& \sum_{k=1}^{n}\mathbb{E}\left(\Pr\left(\mathcal{M}_k^2 > \frac{L}{\mathrm{tr}(\mathbf{G})}\|\psi_k\|_m^2 \middle| \mathbf{y}\right) \middle| \Omega_\delta^m\right) \\
\leq& \sum_{k=1}^{n}\mathbb{E}_+\left(2\exp\left\{-\frac{Lm}{2\mathrm{tr}(\mathbf{G})\widetilde{\eta}_{\max}^2}\right\}\right)\Pr(\Omega_\delta^m)^{-1} \\
=& 2n\mathbb{E}_+\left(\exp\left\{-\frac{Lm}{2\mathrm{tr}(\mathbf{G})\widetilde{\eta}_{\max}^2}\right\}\right)\Pr(\Omega_\delta^m)^{-1} \\
\leq& 2n\exp\left\{-\frac{Lm}{2(1+\delta)n\widetilde{\eta}_{\max}^2}\right\}. \qquad (73)
\end{aligned}$$

Taking now

$$L = 2(1+r)\widetilde{\eta}_{\max}^2(1+\delta)\frac{n\ln m}{m}$$

we obtain

$$\Pr\left(\|\mathbf{J}^\top\widetilde{\boldsymbol{\eta}}\|_2^2 > 2(1+r)(1+\delta)\frac{\widetilde{\eta}_{\max}^2 n\ln m}{m} \middle| \Omega_\delta^m\right) \leq 2n\exp\{-(1+r)\ln m\} \leq 2nm^{-(1+r)}$$

that gives the thesis since $n \leq m$. □

6 Conclusions

We have proven convergence estimates in probability and in expectation for discrete least squares with noisy evaluations at random points. These estimates clarify how the overall approximation error depends on the best approximation error, on the noise terms and on the confidence level. Several noise models have been considered, with different assumptions on the boundedness of the noise. Our analysis quantifies the precise condition between the number of pointwise evaluations, the requested confidence level and the dimension of the underlying approximation space that ensures a stable and accurate discrete least-squares approximation in presence of noise. Finally we have applied our theoretical findings to the particular setting of multivariate polynomial approximation, using results achieved in previous analyses.

7 Acknowledgments

F. Nobile and G. Migliorati acknowledge the support of the Center for ADvanced MOdeling Science (CADMOS). R. Tempone is a member of the KAUST SRI Center for Uncertainty Quantification in Computational Science and Engineering.

References

[1] R.Ahlswede, A.Winter: *Strong converse for identification via quantum channels*, IEEE Trans. Inf. Theory, 48:569–579, 2002.

[2] P.Binev, A.Cohen, W.Dahmen, R.DeVore, V.Temlyakov: *Universal algorithms for learning theory - part I: piecewise constant functions*, J. Mach. Learn. Res., 6:1297–1321, 2005.

[3] P.Binev, A.Cohen, W.Dahmen, R.DeVore: *Universal algorithms for learning theory - part II: piecewise polynomial functions*, Constr. Approx., 26:127–152, 2007.

[4] A.Chkifa, A.Cohen, G.Migliorati, F.Nobile, R.Tempone: *Discrete least squares polynomial approximation with random evaluations - application to parametric and stochastic elliptic PDEs*, ESAIM Math. Model. Numer. Anal., 49(3):815–837, 2015.

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

[6] F.Cucker, S.Smale: *On the mathematical foundations of learning*, Bulletin of the American Mathematical Society, 39(1):1–49, 2001.

[7] F.Cucker, D.Zhou: *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, 2007.

[8] A.Dembo, O.Zeitouni: *Large Deviations Techniques and Applications*, Springer, 1998.

[9] R.G.Ghanem, P.D.Spanos: *Stochastic finite elements: a spectral approach*, Springer-Verlag, New York, 1991.

[10] L.Györfi, M.Kohler, A.Krzyzak, H.Walk: *A distribution-free theory of nonparametric regression*, Springer-Verlag, 2002.

[11] O.Le Maitre, O.M.Knio: *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*, Springer, 2010.

[12] G.Migliorati: *Polynomial approximation by means of the random discrete L2 projection and application to inverse problems for PDEs with stochastic data*, Ph.D. thesis, Dipartimento di Matematica "Francesco Brioschi", Politecnico di Milano, Milano, Italy, and Centre de Mathématiques Appliquées, École Polytechnique, Palaiseau, France, 2013.

[13] G.Migliorati: *Multivariate Markov-type and Nikolskii-type inequalities for polynomials associated with downward closed multi-index sets*, J. Approx. Theory, 189:137–159, 2015.

[14] G.Migliorati, F.Nobile: *Analysis of discrete least squares on multivariate polynomial spaces with evaluations at low-discrepancy point sets*, J. Complexity, 31(4):517–542, 2015.

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

[16] G.Migliorati, F.Nobile, E.von Schwerin, R.Tempone: *Approximation of Quantities of Interest in stochastic PDEs by the random discrete L^2 projection on polynomial spaces*, SIAM J. Sci. Comput., 35(3):A1440–A1460, 2013.

[17] P.Niyogi: *The Informational Complexity of Learning*, Kluwer, 1998.

[18] T.Poggio, S.Smale: *The mathematics of learning: Dealing with data*. Notices Amer.Math.Soc., 50:537–544, 2003.

[19] V.N.Temlyakov: *Approximation in Learning Theory*, Constr.Approx., 27:33–74, 2008.

[20] J.Tropp: *User-Friendly Tail Bounds for Sums of Random Matrices*, Found. Comput. Math., 12:389–434, 2012.

[21] S.R.S.Varadhan: *Special invited paper: Large Deviations*, The Annals of Probability, 36(2):397–419, 2008.

[22] V.Vapnik: *Statistical learning theory*, John Wiley & Sons, 1998.

[23] T.Zhou, A.Narayan, Z.Xu: *Multivariate discrete least-squares approximations with a new type of collocation grid*, SIAM J. Sci. Comput., 36(5), pp. A2401-A2422, 2014.