Journal of Complexity 60 (2020) 101484

Contents lists available at ScienceDirect

Journal of Complexity

journal homepage: www.elsevier.com/locate/jco

On the worst-case error of least squares algorithms for L_2-approximation with high probability[☆]

Mario Ullrich

Institut für Analysis, Johannes Kepler Universität Linz, Austria
Moscow Center for Fundamental and Applied Mathematics, Lomonosov Moscow State University, Russia

ARTICLE INFO

Article history:
Received 25 March 2020
Received in revised form 10 April 2020
Accepted 10 April 2020
Available online 18 April 2020

Keywords:
L_2-approximation
Least squares
Random matrices

ABSTRACT

It was recently shown by D. Krieg and M. Ullrich that, for L_2-approximation of functions from a reproducing kernel Hilbert space, function values are almost as powerful as arbitrary linear information if the approximation numbers are square-summable. That is,

$$e_n \lesssim \sqrt{\frac{1}{k_n}\sum_{j\geq k_n} a_j^2} \qquad \text{with} \qquad k_n \asymp \frac{n}{\ln(n)},$$

where e_n are the sampling numbers and a_k are the approximation numbers. In particular, if $(a_k) \in \ell_2$, then e_n and a_n are of the same polynomial order. For this, we presented an explicit (weighted least squares) algorithm based on i.i.d. random points and proved that this works with positive probability. This implies the existence of a good deterministic sampling algorithm.

Here, we present a modification of our proof that shows that the same algorithm works with probability at least $1 - n^{-c}$ for any given $c > 0$.

Let H be a Hilbert space of real- or complex-valued functions on a set D such that point evaluation

$$\delta_x : H \to \mathbb{K}, \quad f \mapsto f(x),$$

with $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, is a continuous functional for all $x \in D$, which is usually called *reproducing kernel Hilbert space*. We consider numerical approximation of functions from such spaces, using only

[☆] Communicated by E. Novak.

E-mail address: mario.ullrich@jku.at.

https://doi.org/10.1016/j.jco.2020.101484
0885-064X/

function values. We measure the error in the space $L_2 = L_2(D, \mathcal{A}, \mu)$ of square-integrable functions with respect to an arbitrary measure μ such that H is embedded into L_2. This means that H consists of square-integrable functions such that two functions that are equal μ-almost everywhere are also equal pointwise.

We are interested in the *nth minimal worst-case error*

$$e_n := e_n(H) := \inf_{\substack{x_1,\dots,x_n\in D\\ \varphi_1,\dots,\varphi_n\in L_2}} \sup_{f\in H:\|f\|_H\le 1} \Big\| f - \sum_{i=1}^n f(x_i)\,\varphi_i \Big\|_{L_2},$$

which is the worst-case error of an optimal algorithm that uses at most n function values. These numbers are sometimes called *sampling numbers*. We want to compare e_n with the *nth approximation number*

$$a_n := a_n(H) := \inf_{\substack{L_1,\dots,L_n\in H'\\ \varphi_1,\dots,\varphi_n\in L_2}} \sup_{f\in H:\|f\|_H\le 1} \Big\| f - \sum_{i=1}^n L_i(f)\,\varphi_i \Big\|_{L_2},$$

where H' is the space of all bounded, linear functionals on H. This is the worst-case error of an optimal algorithm that uses n linear functionals as information, and it is known that it equals the nth singular value of the embedding $\mathrm{id}: H \to L_2$, see, e.g., [6, Corollary 4.12]. For an exposition of such approximation problems we refer to [6–8], especially [8, Chapter 26 &29], and references therein. The main result of [4] is stated as follows.

Theorem 1 ([4, Theorem 1]). *There are absolute constants $C, c > 0$ and a sequence of natural numbers (k_n) with $k_n \ge cn/\ln(n+1)$ such that the following holds. For any $n \in \mathbb{N}$, any measure space $(D, \mathcal{A}, \mu)$ and any reproducing kernel Hilbert space H of real-valued functions on D that is embedded into $L_2(D, \mathcal{A}, \mu)$, we have*

$$e_n(H)^2 \le \frac{C}{k_n} \sum_{j\ge k_n} a_j(H)^2.$$

We refer to [4] for a thematic classification, further literature, and the implications for some long-standing open problems in the field.

This theorem was extended in [3] to (complex) Hilbert spaces that may not be embedded into L_2, which may happen, e.g., if the support of μ is not equal to D. Moreover, we have learned from [3] about the more recent paper [9], which allows for bounds on the singular values of random matrices with explicit constants. Combining [9] with the proof technique from [4], we do not only see explicit constants from [3]. We obtain that the method described below works with high probability, i.e., with probability at least $1 - n^{-c}$ for all $c > 0$.

Before we state the main result, let us recall the method from [4]:

First of all, let $\mathrm{id}: H \to L_2$ be the embedding from H to $L_2(D, \mathcal{A}, \mu)$ and $W := \mathrm{id}^*\mathrm{id}$. Since W is positive and compact, it follows from the spectral theorem that there is an orthogonal basis $\mathcal{B} = \{b_k : k \in \mathbb{N}\}$ of H that consists of eigenfunctions of W. Without loss of generality, we may assume that H is infinite-dimensional. It is easy to verify that $\mathcal{B}$ is also orthogonal in L_2. We may assume that the eigenfunctions are normalized in L_2 and that $\|b_1\|_H \le \|b_2\|_H \le \dots$, such that $a_k(H) = \|b_{k+1}\|_H^{-1}$.

Now let $k \in \mathbb{N}$ (to be specified later), $x_1, \dots, x_n \in D$ be some given sampling nodes, and $V_k := \mathrm{span}\{b_1, \dots, b_k\}$. We then consider the algorithm $A_{n,k}: H \to L_2$ given by

$$A_{n,k}(f) := \underset{g\in V_k}{\mathrm{argmin}} \sum_{i=1}^n \frac{|g(x_i) - f(x_i)|^2}{\rho_k(x_i)}, \tag{1}$$

where ρ_k is given by

$$\rho_k: D \to \mathbb{R}, \quad \rho_k(x) = \frac{1}{2}\left(\frac{1}{k}\sum_{j<k} b_{j+1}(x)^2 + \frac{1}{\sum_{j\ge k} a_j^2} \sum_{j\ge k} a_j^2 b_{j+1}(x)^2 \right).$$

Note that, under mild assumptions, $A_{n,k}(f) = f$ whenever $f \in V_k$, see (2).

The *worst-case error* of $A_{n,k}$ is defined as

$$e(A_{n,k}, H) := \sup_{f\in H:\|f\|_H\le 1} \left\| f - A_{n,k}(f) \right\|_{L_2},$$

and we have $e_n(H) \le e(A_{n,k}, H)$ for every choice of k and $x_1, \dots, x_n$.

In [4] we proved that, if $x_1, \dots, x_n$ are i.i.d. random points with μ-density ρ_{k_n}, with $k_n \asymp n/\log(n)$, then $e(A_{n,k_n}, H)$ satisfies the bound in Theorem 1 with positive (constant in n) probability. Here, we show that this holds with probability tending to 1, and we determine some explicit constants. (We did not try to optimize them.) Roughly speaking, this shows that, asymptotically, almost all point sets lead to an algorithm for L_2-approximation that satisfies the bound above. This may increase the belief that there is at least some better point set, or even that the conjecture $e_n \asymp a_n$ holds, see, e.g., [8, Open Problem 140].

Our improved result reads as follows.

Theorem 2. *For $n \ge 2$ and $c > 0$, let*

$$k_n := 2 \cdot \left\lfloor \frac{n}{2^8\,(2+c)\,\ln(n)} \right\rfloor.$$

Then, for any measure space $(D, \mathcal{A}, \mu)$ and any reproducing kernel Hilbert space H of real- or complex-valued functions on D that is embedded into $L_2(D, \mathcal{A}, \mu)$, we have

$$e(A_{n,k_n}, H)^2 \le \frac{4}{k_n} \sum_{j\ge k_n/2} a_j(H)^2$$

with probability at least $1 - \frac{8}{n^c}$, where A_{n,k_n} is from (1) and the sampling nodes $x_1, \dots, x_n \in D$ are drawn independently w.r.t. ρ_{k_n}.

The proof

The proof of Theorem 2 is almost the same as given in [4], and is therefore very much inspired by the general technique to assess the quality of *random information* as developed in [1,2]. See also the references collected there. In fact, we only replace [4, Proposition 1] (which is [5, Thm. 2.1]) by [9, Lemma 1] to bound the singular values of the random matrices under consideration.

Let us note that the proof looks rather elementary, and it might be surprising that the results presented in [4] (and here), are not known for some time. However, the way of controlling the 'infinite-dimensional part' by adjusting the density ρ_k accordingly, was seemingly invented in [4], and this turned out to be essential.

First, let $\rho = \rho_k$ and note that the algorithm from (1) can be written as

$$A_{n,k}(f) = \sum_{j=1}^{k} (G^+ Nf)_j b_j, \tag{2}$$

where $N\colon H \to \mathbb{R}^n$ with $N(f) = (\rho(x_i)^{-1/2} f(x_i))_{i\le n}$ is the weighted *information mapping* and $G^+ \in \mathbb{K}^{k\times n}$ is the Moore–Penrose inverse of the matrix

$$G = (\rho(x_i)^{-1/2} b_j(x_i))_{i\le n, j\le k} \in \mathbb{K}^{n\times k}, \tag{3}$$

assuming that G has full rank. In this case, also the argmin in (1) is unique.

To give an upper bound on $e(A_{n,k}, H)$, let us assume that G has full rank. For any $f \in H$ with $\|f\|_H \le 1$, we let $P_k f$ be the orthogonal projection of f onto V_k, and obtain

$$\begin{aligned} \left\| f - A_{n,k}(f) \right\|_{L_2}^2 &\le a_k^2 + \left\| P_k f - A_{n,k}(f) \right\|_{L_2}^2 = a_k^2 + \left\| A_{n,k}(f - P_k f) \right\|_{L_2}^2 \\ &= a_k^2 + \left\| G^+ N(f - P_k f) \right\|_{\ell_2^k}^2 \\ &\le a_k^2 + \left\| G^+\colon \ell_2^n \to \ell_2^k \right\|^2 \left\| N\colon P_k(H)^\perp \to \ell_2^n \right\|^2. \end{aligned}$$

We have used $A_{n,k}(f) \in V_k$ in the first inequality, and $A_{n,k}(P_k f) = P_k f$ in the equality thereafter. The norm of G^+ is the inverse of the kth largest (and therefore the smallest) singular value of the matrix G. The norm of N is the largest singular value of the matrix

$$\Gamma = \big(\rho(x_i)^{-1/2} a_j b_{j+1}(x_i)\big)_{1\leq i\leq n, j\geq k} \in \mathbb{K}^{n\times\infty}. \tag{4}$$

To see this, note that $f = \sum_{j=1}^{\infty} \langle f, b_j\rangle_{L_2} b_j$ converges in H for every $f \in H$, and therefore also pointwise, and that $\|f\|_H^2 = \sum_{j=0}^{\infty} a_j^{-2} |\langle f, b_{j+1}\rangle_{L_2}|^2$. Hence, $N = \Gamma\Delta$ on $P_k(H)^{\perp}$, where the mapping $\Delta \colon P_k(H)^{\perp} \to \ell_2$ with $\Delta f = \Big(\frac{\langle f, b_{j+1}\rangle_{L_2}}{a_j}\Big)_{j\geq k}$ is an isomorphism. This yields

$$e(A_{n,k})^2 \;\leq\; a_k^2 + \frac{s_{\max}(\Gamma)^2}{s_{\min}(G)^2}. \tag{5}$$

It remains to bound $s_{\min}(G)$ from below and $s_{\max}(\Gamma)$ from above. Clearly, any nontrivial lower bound on $s_{\min}(G)$ automatically yields that the matrix G has full rank. To state our results, let

$$\beta_k \;:=\; \Bigg(\frac{1}{k}\sum_{j\geq k} a_j^2\Bigg)^{1/2} \qquad \text{and} \qquad \beta_k' \;:=\; \beta_{\lfloor k/2\rfloor}.$$

Note that $a_{2k}^2 \leq \frac{1}{k}(a_k^2 + \ldots + a_{2k}^2) \leq \beta_k^2$ for all k and thus $\max\{a_k, \beta_k\} \leq \beta_k'$.

The rest of the paper is devoted to the proof of the following two claims:

For each $k \leq \frac{n}{128\cdot(2+c)\cdot\log(n)}$, we have

$$\textbf{Claim 1:} \quad \mathbb{P}\Big(s_{\max}(\Gamma)^2 \;\leq\; n\,\frac{3(\beta_k')^2}{2}\Big) \;\geq\; 1 - \frac{4}{n^c}$$

$$\textbf{Claim 2:} \quad \mathbb{P}\Big(s_{\min}(G)^2 \;\geq\; \frac{n}{2}\Big) \;\geq\; 1 - \frac{4}{n^c}$$

Together with (5) and a union bound, this yields

$$e(A_{n,k})^2 \;\leq\; a_k^2 + 3(\beta_k')^2 \;\leq\; 4\beta_{\lfloor k/2\rfloor}^2$$

with probability at least $1 - \frac{8}{n^c}$, which is the statement of Theorem 2.

Both claims are based on [9, Lemma 1], which we state here in a special case, i.e., we set $\delta = \frac{4}{n^c}$, see [9, top of p. 205]. By $\|M\|$ we denote the spectral norm of M.

Proposition 1. *Let X be a random vector in $\mathbb{C}^k$ or ℓ_2 with $\|X\|_2 \leq R$ with probability 1, and let $X_1, X_2, \ldots$ be independent copies of X. Additionally, let $E := \mathbb{E}(XX^*)$ satisfy $\|E\| \leq 1$, and define, for $c > 0$,*

$$g(n, R, c) \;:=\; 4R\sqrt{\frac{(2+c)\ln(n)}{n}}.$$

If $g(n, R, c) \leq 2$, then

$$\mathbb{P}\Bigg(\bigg\|\sum_{i=1}^{n} X_i X_i^* - nE\bigg\| \;\leq\; n\cdot g(n, R, c)\Bigg) \;\geq\; 1 - \frac{4}{n^c}.$$

Remark 1. Note that there is a typo in [9, Lemma 1]: It is actually proven that the exponent in the estimate should have t^2 for $t < 1$, and $4t - 4$ for $t \geq 1$, instead of $\min\{t^2, 4t - 4\}$. However, the calculations at [9, top of p. 205] are not affected. This issue was also addressed in [3].

Proof of Claim 1. Consider independent copies $X_1, \ldots, X_n$ of the vector

$$X \;=\; \frac{1}{\beta_k'\sqrt{\rho(x)}}\Big(a_k b_{k+1}(x),\, a_{k+1} b_{k+2}(x),\, \ldots\Big)^{\top},$$

where x is a random variable on D with density ρ. Clearly, $\sum_{i=1}^{n} X_i X_i^* = \frac{1}{(\beta_k')^2}\Gamma^*\Gamma$ with Γ from (4). First observe

$$\|X\|_2^2 \;=\; \frac{1}{(\beta_k')^2\rho(x)}\sum_{j\geq k} a_j^2\, b_{j+1}(x)^2 \;\leq\; \frac{2}{(\beta_k')^2}\sum_{j\geq k} a_j^2 \;=\; 2k \;=:\; R^2.$$

Since $E=\mathbb{E}(XX^*)=\mathrm{diag}(\frac{a_k^2}{(\beta_k')^2},\frac{a_{k+1}^2}{(\beta_k')^2},\dots)$, we have $\|E\|=\frac{a_k^2}{(\beta_k')^2}\leq 1$.

Using $k\leq\frac{n}{128(2+c)\ln(n)}$ and

$$g(n,R,c) \;=\; g\left(n,\sqrt{2k},c\right) \;=\; \sqrt{32\,(2+c)\,k\frac{\ln(n)}{n}} \;\leq\; \frac{1}{2},$$

we obtain from Proposition 1 that

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} X_i X_i^* - nE\right\| \leq \frac{n}{2}\right) \geq 1-\frac{4}{n^c}.$$

This implies

$$\begin{aligned} s_{\max}(\Gamma)^2 &= \left\|\Gamma^*\Gamma\right\| = (\beta_k')^2\left\|\sum_{i=1}^{n} X_i X_i^*\right\| \leq (\beta_k')^2\left(\|nE\| + \left\|\sum_{i=1}^{n} X_i X_i^* - nE\right\|\right)\\ &\leq n\,a_k^2 + n\frac{(\beta_k')^2}{2} \end{aligned}$$

with probability at least $1-4/n^c$ for all $k\leq\frac{n}{128(2+c)\ln(n)}$. This yields Claim 1. □

Proof of Claim 2. Consider $X=\rho(x)^{-1/2}(b_1(x),\dots,b_k(x))^\top$ with x distributed according to ρ. Clearly, $\sum_{i=1}^{n} X_i X_i^* = G^*G$ with G from (3). First observe

$$\|X\|_2^2 \;=\; \rho(x)^{-1}\sum_{j\leq k} b_j(x)^2 \;\leq\; 2k \;=:\; R^2.$$

Since $E=\mathbb{E}(XX^*)=\mathrm{diag}(1,\dots,1)$, we have $\|E\|=1$.

Again, for $k\leq\frac{n}{128(2+c)\ln(n)}$, we obtain $g(n,R,c)\leq\frac{1}{2}$, and therefore, from Proposition 1, that $\mathbb{P}\left(\left\|\sum_{i=1}^{n} X_i X_i^* - nE\right\|\leq\frac{n}{2}\right)\geq 1-\frac{4}{n^c}$. This implies that

$$s_{\min}(G)^2 = s_{\min}(G^*G) \geq s_{\min}(nE) - \|G^*G-nE\| \geq n/2$$

with probability at least $1-\frac{4}{n^c}$, and yields Claim 2. □

Acknowledgment

I thank David Krieg for the indispensable discussions. He decided not to be a co-author, because he thinks that he did not contribute enough.

References

[1] A. Hinrichs, D. Krieg, E. Novak, J. Prochno, M. Ullrich, On the power of random information, in: Radon Series on Computational and Applied Mathematics, Vol. 27, 2019, (in press) arXiv:1903.00681.

[2] A. Hinrichs, D. Krieg, E. Novak, J. Prochno, M. Ullrich, Random sections of ellipsoids and the power of random information, 2019, arXiv:1901.06639.

[3] L. Kämmerer, T. Ullrich, T. Volkmer, Worst case recovery guarantees for least squares approximation using random samples, 2019, arXiv:1911.10111.

[4] D. Krieg, M. Ullrich, Function values are enough for L_2-approximation, 2019, arXiv:1905.02516.

[5] S. Mendelson, A. Pajor, On singular values of matrices with independent rows, Bernoulli 12 (5) (2006) 761–773.

[6] E. Novak, H. Woźniakowski, Tractability of Multivariate Problems. Vol. 1: Linear Information, in: EMS Tracts in Mathematics, vol. 6, European Mathematical Society (EMS), Zürich, 2008.

[7] E. Novak, H. Woźniakowski, Tractability of Multivariate Problems. Volume II: Standard Information for Functionals, in: EMS Tracts in Mathematics, vol. 12, European Mathematical Society (EMS), Zürich, 2010.

[8] E. Novak, H. Woźniakowski, Tractability of Multivariate Problems. Volume III: Standard Information for Operators, in: EMS Tracts in Mathematics, vol. 18, European Mathematical Society (EMS), Zürich, 2012.

[9] R.I. Oliveira, Sums of random Hermitian matrices and an inequality by Rudelson, Electron. Commun. Probab. 15 (2010) 203–212.