ON THE WORST-CASE ERROR OF LEAST SQUARES ALGORITHMS FOR \( L_2 \)-APPROXIMATION WITH HIGH PROBABILITY

MARIO ULLRICH

\( \text{Institut für Analysis, Johannes Kepler Universität Linz, Austria} \)

Abstract. It was recently shown in [4] that, for \( L_2 \)-approximation of functions from a Hilbert space, function values are almost as powerful as arbitrary linear information, if the approximation numbers are square-summable. That is, we showed that

\[
\varepsilon_n \lesssim \sqrt{\frac{1}{k_n} \sum_{j \geq k_n} a_j^2} \quad \text{with} \quad k_n \approx \frac{n}{\ln(n)},
\]

where \( \varepsilon_n \) are the sampling numbers and \( a_k \) are the approximation numbers. In particular, if \((a_k) \in \ell_2\), then \( \varepsilon_n \) 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 the proof in [4] that shows that the same algorithm works with probability at least \( 1 - n^{-c} \) for all \( c > 0 \).

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

\[
\delta_x : H \rightarrow \mathbb{R}, \quad f \mapsto f(x)
\]

is a continuous functional for all \( x \in D \), which are usually called \textit{reproducing kernel Hilbert spaces}. We consider numerical approximation of functions from such spaces, using only 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 \( \mu \) such that \( H \) is embedded into \( L_2 \). This means that \( H \) consists of square-integrable functions such that two functions that are equal \( \mu \)-almost everywhere are also equal point-wise.

\textit{E-mail address:} mario.ullrich@jku.at

\textit{Date:} January 31, 2022.

2010 \textit{Mathematics Subject Classification.} 41A25, 41A46, 60B20;

\textit{Key words and phrases.} \( L_2 \)-approximation, least squares, random matrices.
We are interested in the $n$-th minimal worst-case error 

$$e_n := e_n(H) := \inf_{x_1, \ldots, x_n \in D} \sup_{\varphi_1, \ldots, \varphi_n \in L_2} \left\| f - \sum_{i=1}^n f(x_i) \varphi_i \right\|_{L_2},$$

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 $n$-th approximation number 

$$a_n := a_n(H) := \inf_{L_1, \ldots, L_n \in H'} \sup_{f \in H: \|f\|_H \leq 1} \left\| f - \sum_{i=1}^n L_i(f) \varphi_i \right\|_{L_2},$$

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 $n$-th singular value of the embedding $\text{id}: H \to L_2$. For an exposition of such approximation problems we refer to [6, 7, 8], especially [8, Chapter 26 & 29], and references therein. The main result of [4] is stated as follows.

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

$$e_n(H)^2 \leq \frac{C}{k_n} \sum_{j \geq k_n} a_j(H)^2.$$
generality, we may assume that $H$ is infinite-dimensional. It is easy to verify that $B$ is also orthogonal in $L_2$. We may assume that the eigenfunctions are normalized in $L_2$ and that $\|b_1\|_H \leq \|b_2\|_H \leq \ldots$, such that $a_k(H) = \|b_{k+1}\|_H^{-1}$.

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

$$A_{n,k}(f) := \arg\min_{g \in V_k} \sum_{i=1}^n \frac{|g(x_i) - f(x_i)|^2}{g_k(x_i)},$$

where $\rho = \rho_k$ is given by

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

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

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

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

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

In [4] we proved that, if $x_1, \ldots, x_n$ are i.i.d. random points with $\mu$-density $\rho$, then $e(A_{n,k_n})$ with $k_n \asymp n/\log(n)$ satisfies the bound in Theorem 1 with positive 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 in the conjecture $e_n \asymp a_n$, see e.g. [8, Open Problem 140]

Our improved result reads as follows.

**Theorem 2.** For $n \geq 2$ 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

$$e_n(A_n, H)^2 \leq \frac{4}{k_n} \sum_{j \geq k_n/2} a_j(H)^2$$

with probability at least $1 - \frac{8}{n^c}$, where $A_n = A_{n,k_n}$ from (1).
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 \( \rho \) accordingly, was seemingly invented in [4], and this turned out to be essential.

First, let us note that the algorithm from (1) can be written as

\[
A_{n,k}(f) = \sum_{j=1}^{k} (G^+ N f) j b_j,
\]

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

\[
G = (\rho(x_i)^{-1/2} b_j(x_i))_{i \leq n, j \leq k} \in \mathbb{R}^{n \times k},
\]

assuming that \( G \) has full rank.

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

\[
\|f - A_{n,k}(f)\|_{L^2}^2 \leq a_k^2 + \|P_k f - A_{n,k}(f)\|_{L^2}^2 = a_k^2 + \|A_{n,k}(f - P_k f)\|_{L^2}^2
\]
\[
= a_k^2 + \|G^+ N(f - P_k f)\|_{\ell_2^k}^2
\]
\[
\leq a_k^2 + \|G^+ : \ell_2^n \to \ell_2^k\|_{\ell_2^k}^2 \|N : P_k(H) \perp \to \ell_2^n\|_{\ell_2^n}^2.
\]

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

\[
\Gamma = (\rho(x_i)^{-1/2} a_j b_{j+1}(x_i))_{1 \leq i \leq n, j \geq k} \in \mathbb{R}^{n \times \infty}.
\]
To see this, note that 
\[ f = \sum_{j=1}^{\infty} (f, b_j)_{L^2} b_j \] converges in \( H \) for every \( f \in H \), and therefore also point-wise, and that \( \|f\|^2_H = \sum_{j=0}^{\infty} a_j^{-2} |(f, b_{j+1})_{L^2}|^2 \). Hence, \( N = \Gamma \Delta \) on \( P_k(H)^\perp \), where the mapping \( \Delta : P_k(H)^\perp \rightarrow \ell^2 \) with \( \Delta f = \left( \frac{(f, b_{j+1})_{L^2}}{a_j} \right)_{j \geq k} \) is an isomorphism. This yields

\[ e(A_{n,k})^2 \leq a_k^2 + \frac{s_{\max}(\Gamma)^2}{s_{\min}(G)^2}. \]

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

\[ \beta_k := \left( \frac{1}{k} \sum_{j \geq k} a_j^2 \right)^{1/2} \quad \text{and} \quad \beta'_k := \beta_{\lfloor k/2 \rfloor}. \]

Note that \( a_k^2 \leq \frac{1}{k} (a_k^2 + \ldots + a_{2k}^2) \leq \beta_k^2 \) 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 \( k \leq \frac{n}{128(2+c) \cdot \log(n)} \), we have

**Claim 1:** \( \mathbb{P} \left( s_{\max}(\Gamma)^2 \leq \frac{3(\beta'_k)^2}{2} \right) \geq 1 - \frac{4}{n^c} \)

**Claim 2:** \( \mathbb{P} \left( s_{\min}(G)^2 \geq \frac{n}{2} \right) \geq 1 - \frac{4}{n^c} \)

Together with (2) 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 \( \ell^2 \) 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

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

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

\[ \mathbb{P} \left( \left\| \sum_{i=1}^{n} X_i X_i^* - nE \right\| \leq n \cdot g(n, R, c) \right) \geq 1 - \frac{4}{n^c}. \]
Proof of Claim 1. Consider independent copies $X_1, \ldots, X_n$ of the vector

$$X = \frac{1}{\beta_k} \sqrt{g(x)} (a_kb_{k+1}(x), a_{k+1}b_{k+2}(x), \ldots) \top,$$

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

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

Since $E = \mathbb{E}(XX^*) = \text{diag}(\frac{a_1^2}{(\beta_1)^2}, \frac{a_2^2}{(\beta_2)^2}, \ldots)$, we have $\|E\| = \frac{a_k^2}{(\beta_k)^2} \leq 1$. Using $k \leq \frac{n}{128(2+c) \ln(n)}$ and

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

we obtain from Proposition 1 that

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

This implies

$$s_{\max}(\Gamma)^2 = \|\Gamma^* \Gamma\| = (\beta_k')^2 \left\| \sum_{i=1}^n X_iX_i^* \right\| \leq (\beta_k')^2 \left( \|nE\| + \left\| \sum_{i=1}^n X_iX_i^* - nE \right\| \right)$$

$$\leq n a_k^2 + n \left( \frac{\beta_k'}{2} \right)^2$$

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

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

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

Since $E = \mathbb{E}(XX^*) = \text{diag}(1, \ldots, 1)$ 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 $\mathbb{P}\left( \left\| \sum_{i=1}^n X_iX_i^* - nE \right\| \leq \frac{n}{2} \right) \geq 1 - \frac{4}{n^c}$. This implies that

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

with probability at least $1 - \frac{4}{n^c}$, and yields Claim 2. \qed
Acknowledgement. I thank David Krieg for the indispensable discussions. He decided not to be a co-author, because he thinks that he didn’t contribute.

References

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

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

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

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

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

[6] E. Novak and H. Woźniakowski. *Tractability of multivariate problems. Vol. 1: Linear information*, volume 6 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2008.

[7] E. Novak and H. Woźniakowski. *Tractability of multivariate problems. Volume II: Standard information for functionals*, volume 12 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2010.

[8] E. Novak and H. Woźniakowski. *Tractability of multivariate problems. Volume III: Standard information for operators*, volume 18 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2012.

[9] R. I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electr. Comm. Probab.*, 15:203–212, 2010.