Journal of Applied and Numerical Analysis
2023, Vol. 1, pp. 88–126
doi: 10.30970/ana.2023.1.88

On the power of iid information for linear approximation

M. Sonnleitner[1], M. Ullrich[2]

[1]*Faculty of Computer Science and Mathematics, University of Passau, 94032 Passau, Germany,*
[2]*Institute of Analysis, Johannes Kepler University Linz, 4040 Linz, Austria*

Анотація. В роботі розглянуто силу випадкової інформації для наближення в (детерміністичному) найгіршому випадку, з особливим акцентом на інформацію, що складається з вибраних функціоналів незалежно та однаково розподілених (iid) випадковим чином на класі допустимих інформаційних функціоналів. Ми представляємо загальний результат на основі методу зважених найменших квадратів та наслідки для особливих випадків. Покращення доступні, якщо інформація є "гаусівською" або якщо ми розглядаємо значення функції iid для просторів Соболєва. Ми включили відкриті запитання, щоб спрямувати майбутні дослідження про силу випадкової інформації в контексті інформаційної складності.

Abstract. This survey is concerned with the power of random information for approximation in the (deterministic) worst-case setting, with special emphasis on information consisting of functionals selected independently and identically distributed (iid) at random on a class of admissible information functionals. We present a general result based on a weighted least squares method and derive consequences for special cases. Improvements are available if the information is "Gaussian" or if we consider iid function values for Sobolev spaces. We include open questions to guide future research on the power of random information in the context of information-based complexity.

Contents

1 Introduction

This survey is oriented towards *information-based complexity* (IBC) and we refer to [90, 95, 97, 113] for a more comprehensive treatment of information-based complexity. For an introduction to

2020 *Mathematics Subject Classification*: 65-02, 41A25, 47B06, 68Q25, 94A20.
Key words: information-based complexity, random information, least squares, random matrices.

the related field of optimal recovery we refer to [26,34,102,111]. Perhaps the most prominent display of the power of iid information, as we understand it, is the field of compressed sensing (or sparse recovery) which is presented in [30] from the viewpoint of IBC. We will only briefly touch upon this direction as our focus is on *linear approximation*.

Numerical approximation, as considered here, is formally specified by two normed real vector spaces, say H and G, of functions on a set D, a subset $F \subset H$ for which also $F \subset G$, and a class of *admissible information* Λ consisting of functionals on F. (We often consider F to be the unit ball B_H in H and assume that the functionals are defined on H.) The aim is to approximate the *a priori unknown* $f \in F$ based on n pieces of information (or measurements) $\ell_1(f), \dots, \ell_n(f)$ with $\ell_i \in \Lambda$ such that we can guarantee a small error with respect to the norm in G. In general, one does not have access to arbitrary measurements, which is the reason for restricting to Λ. Typical examples for admissible information are

- $\Lambda^{\mathrm{all}} := H'$, i.e., all continuous linear functionals on H,
- $\Lambda =$ "certain expectations of the input function",
- $\Lambda =$ "coefficients w.r.t. a given basis, wavelets etc." or
- $\Lambda^{\mathrm{std}} := \{\delta_x \colon \delta_x(f) = f(x),\ f \in H,\ x \in D\}$ (*function values*).

It is desirable to minimize the approximation error, which is achieved by the "best information" from a given class Λ. To make this precise, we identify information with an *information mapping* of the form

$$N_n \colon H \to \mathbb{R}^n, \quad N_n(f) = \big(\ell_1(f), \dots, \ell_n(f)\big), \quad f \in H, \tag{1.1}$$

with $\ell_1, \dots, \ell_n \in \Lambda$. (One may also consider adaptive information, i.e., ℓ_j may depend on the already computed $\ell_1(f), \dots, \ell_{j-1}(f)$, but we do not treat this here.) Any approximation method (or *algorithm*) based on the information N_n will be of the form

$$A_n(f) \,=\, \varphi_n \circ N_n(f) \,=\, \varphi_n\big(\ell_1(f), \dots, \ell_n(f)\big), \quad f \in H, \tag{1.2}$$

where $\varphi_n \colon \mathbb{R}^n \to G$ is an arbitrary mapping. Linear approximation is concerned with the case of linear $A_n \colon H \to G$. The *worst-case error (w.c.e.)* of an algorithm A_n as in (1.2) is then defined by

$$e(A_n, F, G) \,:=\, \sup_{f \in F} \|f - A_n(f)\|_G$$

and any upper bound on $e(A_n, F, G)$ guarantees an a priori error bound on A_n in the class F. Such a bound should be compared to the best possible error bounds for the given (class of) information.

First, if an information mapping N_n as in (1.1) is fixed, define the *radius of information* N_n by

$$r(N_n, F, G) \,:=\, \inf_{\varphi_n \colon \mathbb{R}^n \to G} \sup_{f \in F} \Big\| f - \varphi_n \circ N_n(f) \Big\|_G \tag{1.3}$$

which quantifies the quality or *power* of N_n. This should be seen relative to the "best information" from the class Λ which gives rise to the *n-th minimal error of information from* Λ defined by

$$e_n(F, G, \Lambda) \,:=\, \inf_{N_n \in \Lambda^n} r(N_n, F, G) \,=\, \inf_{A_n} e(A_n, F, G),$$

where the latter infimum is over all algorithms of the form (1.2) with $\ell_1, \dots, \ell_n \in \Lambda$.

The number $e_n(F, G, \Lambda)$ (or rather the associated sequence) quantifies the power of optimal information and serves as the benchmark for any information obtainable from Λ. With this benchmark at our disposal, we clarify what "iid" information is.

Independently and identically distributed (iid) information is given by independent random continuous functionals $\ell_1, \dots, \ell_n \in \Lambda$, either on F or H with respect to a metric or norm, defined on some probability space $(\Omega, \Sigma, \mathbb{P})$ and having a common distribution ν on Λ. It is necessarily

nonadaptive. A random measurement of $f \in F$ is given by a random variable $\ell_j(f)$, i.e., a realization $\ell_j(f)^\omega$, $\omega \in \Omega$, of $\ell_j(f)$ is the application of the realization $\ell_j^\omega \in \Lambda$ of ℓ_j to f. Thus, $N_n^{(\nu)}(f) = (\ell_1(f), \dots, \ell_n(f))$ is a random vector with distribution $f_*(\nu)^{\otimes n}$ on $\mathbb{R}^n$, where $f_*(\nu)$ is the pushforward measure under $\ell \mapsto \ell(f)$. (We suppress the ω and the corresponding probability space in the following.)

For each realization of $N_n^{(\nu)}$ we study again the minimal error that can be achieved with this information. That is, we consider the random variable

$$e_n^{iid}(F, G, \nu) \;:=\; r(N_n^{(\nu)}, F, G),$$

which we call the *n-th minimal error of iid information* w.r.t. ν from Λ, see also (1.3). It is clearly of interest to study characteristics of the above random variable, but it is still not clear what a reasonable quantity for this is. We use for example bounds holding in expectation or with high probability (whp), i.e., with probability tending to one as n goes to infinity. In any case, we ignore events of measure zero.

The distribution or probability measure ν depends on the problem. To illustrate this, let us discuss some examples which are also our main applications.

First, we consider *standard information*, i.e., if $\Lambda = \Lambda^{\mathrm{std}}$, for $F \subset L_2(\mu)$ with some measure μ on D. (Additional assumptions will guarantee that point evaluations are well-defined.) A distribution on this class of information corresponds to a distribution ν on the domain D, if we consider $\ell(f) = f(X)$ where X has distribution ν on D. A natural choice of distribution is given by $\nu = \mu$, but we sometimes need another distribution for proving "near-optimal" results.

On the class of arbitrary linear functionals we consider a Gaussian measure giving rise to *Gaussian information*, i.e. Gaussian random functionals. Employing an interpretation from the finite-dimensional setting, this corresponds to the radius of the intersection of F with a random subspace, a classical problem from geometric analysis and Banach space theory, see Section 4.3.

Finally, let us also mention *random Fourier coefficients*, which are given by $\ell(f) = \langle f, b_K \rangle$ with a fixed orthonormal basis $\{b_k\}_{k \in \mathbb{N}}$ of H, where K has distribution ν on $\mathbb{N}$ and H is a Hilbert space. We will present upper and lower bounds in all these cases.

As it is our main object of study, we will use "random information" mostly synonymously with "iid information" and thus frequently speak of the power of random information. To summarize, we study $e_n^{iid}(F, G, \nu)$ for

- approximation of functions from a class $F \subset H$,
- where the error is measured in a normed space G with $F \subset G$, and
- with a *random information mapping* $N_n^{(\nu)} \colon F \to \mathbb{R}^n$,
- where ν is a probability measure on the class of admissible information Λ.

We state several reasons for studying iid information in this setting:

1. If $e_n^{iid}(F, G, \nu)$ is "small" with positive probability, we get an upper bound for $e_n(F, G, \Lambda)$, i.e., the error of *optimal approximation*, without the need of finding a sophisticated construction. This is in the spirit of the well-known *probabilistic method*.
2. In cases where the *minimal worst-case error* is known, one might wonder whether the optimal information is somehow special. One approach to this is to study $e_n^{iid}(F, G, \nu)$, and see whether it is with high probability close to optimal, showing that "almost any" information is good, or not, which indicates that a more involved construction is needed.
3. It is a typical assumption in applications, such as machine learning, that information (or data) is given by iid samples with respect to an unknown distribution. It is therefore of interest to identify classes $F \subset G$ and distributions on Λ that allow for reliable error guarantees (such as whp for all $f \in F$) for random information.

4. Further, iid information is often *universal*, that is, useful for many different problems and not particular to a certain problem instance or function class. This is beneficial if the available a priori knowledge is insufficient. In contrast, deterministic constructions of such universal methods are often unknown.

Obviously, the idea of employing randomness for worst-case analysis is not new. However, the power of iid information for linear approximation in a rather general setting, even for restricted classes of information, seems to have been observed only recently. To the best of our knowledge, it was only in the survey [53] that a systematic study was initiated. Since then, there has been some major progress in the case of Gaussian random information and random standard information. In particular, the power of random information has been determined precisely for certain natural choices of F, G and ν, and also some general relations between minimal errors for different classes Λ have been obtained. The present work aims to survey these recent results and to put them into a general framework. Thus, it can be understood as an update to the survey [53].

Let us note that this work builds on the PhD thesis [107] of the first, and the habilitation thesis of the second author which both contain many of the mentioned recent results obtained together with several coauthors.

1.1 What this survey is not about

There are numerous aspects in (optimal) numerical approximation where randomness plays an important role. However, this survey is about aspects that are special to optimal, deterministic approximation based on iid information in the setting described above, and we therefore refrain from discussing indirectly related subjects and results in detail.

The most important omissions we are aware of consist of the following:

- **Nonlinear algorithms.** Random constructions of "good" algorithms played a major role in numerical analysis in the last decades to tackle problems where explicit constructions are not available. Often, this is the case for problems where nonlinear algorithms are required, such as in compressed sensing. We show that sharp results, and some interesting open problems, can be found in the case of linear algorithms, too.
- **Randomized error criterion.** The setting considered in this survey should not be confused with the study of randomized algorithms (sometimes called Monte Carlo methods) with regard to probabilistic bounds on the error for *each* individual f. Here, although we assume that data is produced by random functionals, this information is used for all $f \in F$ *simultaneously* which in general is a stronger error criterion.
- **Other distributions.** In the following, we often assume that the information is iid with respect to a given "optimized" or "natural" distribution. It is clearly of interest, but not our focus, to study the effect of using other distributions or "non-iid" randomness.
- **Implementation and computational cost.** We are not concerned with implementation cost of specific algorithms but focus on information complexity, which is in general only a lower bound on the total computational cost.
- **Adaptive algorithms.** In many cases it is interesting to study *adaptive information/algorithms*. We will only discuss non-adaptive algorithms and note that, if F is convex and symmetric, then the corresponding minimal errors (in the deterministic setting) differ by at most a factor of 2, see [37] and also [95, Section 4.2.1].

The first three topics will be briefly discussed in Section 5 together with additional aspects such as subsampling, learning and tractability of high-dimensional problems. We have to omit many areas where iid information has proven useful, both within and outside of the IBC-framework. These include, for example, density estimation, discretization, numerical integration (Monte Carlo) and Bayesian inference.

Notation: For a measure space (D, Σ, μ), we write $\mathcal{L}_p(\mu)$ for the set of p-integrable functions with usual norm, and inner product for $p = 2$, and denote by $L_p := L_p(\mu)$ the normed space of

corresponding equivalence classes. Whenever convenient, we identify a function with its equivalence class. Moreover, we write $F \hookrightarrow G$ for two metric spaces $F \subset G$ (with possibly different metrics), and say that F *is embedded into* G, if the identity $\mathrm{id}\colon F \to G$, $\mathrm{id}(f) = f$, is a continuous injection. (If G consists of equivalence classes, e.g., for $G = L_p$, then we use the usual modifications.) For two sequences, $(e_n)_{n\geq 0}$ and $(g_n)_{n\geq 0}$, we write $e_n \lesssim g_n$ for $e_n \leq C\, g_n$ for some constant $C > 0$ and all $n \geq 2$, and $e_n \asymp g_n$ if $e_n \lesssim g_n$ and $g_n \lesssim e_n$. If the sequences depend only on certain parameters, say d or s, we write, e.g., $e_n \lesssim_{s,d} g_n$ and $e_n \asymp_s g_n$, respectively, to indicate the dependencies of the hidden constants. Without indication it may depend on all involved parameters, except for n. Given two, possibly infinite, square matrices A and B we indicate the Loewner order by $A \geq B$ meaning that $A - B$ is positive semi-definite. Similarly, we use $A \leq B$ for $B \geq A$. The infinite identity matrix representing the identity $\mathrm{id}\colon \ell_2 \to \ell_2$ is denoted by I and the $n \times n$ identity matrix by I_n.

2 Some benchmarks of optimal approximation

In order to assess the power of random information for numerical approximation, optimal information will serve as a benchmark. Depending on the allowed algorithms or information, this gives rise to different concepts related to the *widths* or *s-numbers* of embeddings. We refer to the monographs [90,100,102] for more information.

Approximation numbers are the minimal errors achievable by an arbitrary linear algorithm. That is, we define the *n-th approximation numbers* (which are sometimes called *linear n-widths*) of $F \subset H$ in G by

$$a_n(F,G) := \inf_{\substack{\ell_1,\dots,\ell_n \in \Lambda^{\mathrm{all}} \\ g_1,\dots,g_n \in G}} \sup_{f\in F} \Big\| f - \sum_{i=1}^{n} \ell_i(f)\, g_i \Big\|_G. \tag{2.1}$$

It is well known that nonlinear algorithms are often superior to linear ones, i.e., $e_n(F,G,\Lambda^{\mathrm{all}}) < a_n(F,G)$ holds. However, equality holds, e.g., if $F = B_H$ is the unit ball of a Hilbert space H, or if F is convex and symmetric and $G = L_\infty$. In such cases, $a_n(F,G) = e_n(F,G,\Lambda^{\mathrm{all}})$, i.e., it is enough to consider linear algorithms, see e.g. [95, Section 4.2]. Moreover, if F is the unit ball B_H in a Banach space H and $H \hookrightarrow G$, then $a_n(F,G) \leq (1+\sqrt{n})\, e_n(F,G,\Lambda^{\mathrm{all}})$, see [95, Theorem 4.9]. In addition, linear algorithms have (practical) advantages, which are not part of this survey. Let us just note that the theory of linear approximations is much more developed than its nonlinear counterpart, with typical techniques such as linear regression, (polynomial) interpolation and projections on certain subspaces.

Kolmogorov widths are another prominent benchmark. The *Kolmogorov n-width* of a set $F \subset G$ is defined by

$$d_n(F,G) := \inf_{\substack{V_n \subset G \\ \dim(V_n)=n}} \sup_{f\in F} \inf_{g\in V_n} \big\| f - g \big\|_G, \tag{2.2}$$

i.e., it is the minimal distance (in G) that is achievable if we were to choose the *best* element from a linear subspace of dimension n.

For this reason, the Kolmogorov widths are in general not related to the theory of algorithms: The inner infimum may not be attained by any (linear) algorithm and it appears to be an "unfair" benchmark.

Still, it is an essential tool in many arguments as it corresponds to the existence of good subspaces which can be used to define algorithms. There are several relations between d_n and a_n, as well as to Gelfand widths via duality theory. For example, we have $d_n(F,L_2) = a_n(F,L_2)$ for $F \subset L_2$ since in this case the best approximation in a subspace is given by orthogonal projection.

Gelfand numbers are closely related to the minimal worst-case errors achievable with arbitrary algorithms based on arbitrary linear information. That is, if we define the *n-th Gelfand number* of

$F \subset H$ in G by

$$c_n(F,G) := \inf_{\substack{W_n \subset H \\ \operatorname{codim}(W_n) \le n}} \sup_{f \in F \cap W_n} \|f\|_G, \tag{2.3}$$

then it is well known that $c_n(F,G)$ differs from $e_n(F,G,\Lambda^{\mathrm{all}})$ by a factor of at most 2 whenever $F \subset H$ is convex and symmetric, see e.g. [95, Section 4.2]. See also [17, 30] for a more general version of this equivalence.

Sampling numbers are the minimal worst-case errors that can be achieved with algorithms based on function values as information, i.e.,

$$g_n(F,G) := e_n(F,G,\Lambda^{\mathrm{std}}) = \inf_{\substack{x_1,\dots,x_n \in D \\ \varphi\colon \mathbb{R}^n \to G}} \sup_{f \in F} \Big\| f - \varphi(f(x_1),\dots,f(x_n)) \Big\|_G, \tag{2.4}$$

where $F \subset G$ are classes of functions on the set D.

The ubiquity of function values in applications might suggest that sampling numbers have been studied in depth as a benchmark. However, besides plenty of results in specific settings, there are only few general results about them. An example is a recent optimal bound for L_2-approximation in Hilbert spaces, which is based on the general results on iid information and subsampling described in Section 5.1, see Theorem 5.1.

Restricting to linear algorithms gives rise to the *linear sampling numbers*

$$g_n^{\mathrm{lin}}(F,G) := \inf_{\substack{x_1,\dots,x_n \in D \\ g_1,\dots,g_n \in G}} \sup_{f \in F} \Big\| f - \sum_{i=1}^n f(x_i)\, g_i \Big\|_G,$$

which can be used to bound $g_n(F,G)$ from above.

Remark 2.1. *We comment on the difference of "width" and "numbers" in the above context. While the width of a set $F \subset G$ can be defined solely based on the knowledge of F and (the norm of) G, the definition of the (approximation/Gelfand) numbers also requires the normed space H containing F to define the class Λ^{all} of all linear functionals. We refer again to [100, 102], and note that there is also a concept of Gelfand width that can yield different results, see [32]. This can also be the case for linear widths/approximation numbers, see [46], and Remark 2.8 in [61].*

3 Approximation based on iid information

In the following we develop a general approach for approximation based on random information and a very simple linear algorithm: A *weighted least squares method* on a suitable subspace and with suitable (explicitly given) weights. This will be done in four steps: We show

1. an L_2-error bound for given information mappings,
2. how this relates to (infinite) matrices in the case of Hilbert spaces,
3. how concentration inequalities lead to optimal bounds, and finally,
4. how we can treat more general classes and approximation w.r.t. other norms.

3.1 L_2-approximation and least squares methods

We first treat the case of approximation in L_2-norm. For this, we fix some measure space (D,Σ,μ), and write $L_p := L_p(\mu)$. We consider the *weighted least squares estimator*

$$A_N(f) := \operatorname*{argmin}_{g \in V_n} \sum_{i=1}^N w_i\, |\ell_i(f-g)|^2 \tag{3.1}$$

for some subspace V_n of L_2 of dimension n, some weights $w_i > 0$ and linear functionals $\ell_i \in \Lambda$, $i = 1, \dots, N$. Note that this map is well-defined and linear under condition (3.2) below, see Proposition 3.1 and its proof. Since we have different types of admissible information Λ in mind, we present results in more general form than in the literature.

In any case, the algorithm A_N in (3.1) is well studied, see e.g. the recent contributions [11,15,18] and the references therein. It seems surprising that, when fed with random information, this simple method often leads to (near-)optimal bounds. The next result forms a basis for many results discussed in this survey, such as the ones in [29,54,72].

Proposition 3.1. *Let $H \subset L_2$ be a normed space of functions on D, and, for $n \le N$, let V_n be a n-dimensional subspace of H, and $\ell_1, \dots, \ell_N$ be linear functionals on H. Assume that*

$$\inf_{g \in V_n} \frac{\sqrt{\sum_{i=1}^N w_i \, |\ell_i(g)|^2}}{\|g\|_{L_2}} \;\ge\; \alpha \tag{3.2}$$

for some $\alpha > 0$ and some weights $w_1, \dots, w_N > 0$. Then, for all $f \in H$ and $g \in V_n$, the algorithm from (3.1) with the corresponding V_n, ℓ_i and w_i is well-defined and satisfies

$$\Big\| f - A_N(f) \Big\|_{L_2} \;\le\; \big\| f - g \big\|_{L_2} + \frac{1}{\alpha} \sqrt{\sum_{i=1}^N w_i \, |\ell_i(f-g)|^2}.$$

Note that, for studying the error over a set $F \subset H$, it is necessary to assume linearity of the used functionals on the affine spaces $f + V_n$ for each $f \in F \cup \{0\}$, and introducing such a surrounding normed space H seems a convenient way to do so.

The above proposition shows that information functionals ℓ_i are "good" for L_2-approximation in F if one can choose a subspace V_n and weights w_i such that the squared sum in (3.2) has large values on V_n and, for each $f \in F$, there is some $g \in V_n$ such that $\|f-g\|_{L_2}$ and $\sum_{i=1}^N w_i \, |\ell_i(f-g)|^2$ are small.

Condition (3.2) says that the discrete (semi)norm based on the functionals and given weights should be comparable to the L_2-norm on V_n. Finding such functionals and weights in the case of standard information is called *discretization* (another interesting topic that we do not discuss in detail). That is, one wants to find a point set such that the L_2-norm of all functions from some n-dimensional V_n can be "discretized" using function values at these points. Also here, random points are used as a tool, and variants of Proposition 3.1 appear. We refer to [23, 44, 63] and references therein.

For approximation in classes of functions, we note that Proposition 3.1, in general, also requires a suitable discretization of the "remainder" $f-g$, which appears to be more involved. This simplifies slightly if we compare the L_2-error of A_N with best approximation in the uniform norm, see e.g. [112, Theorem 2.1]. As usual, set $\|f\|_\infty := \sup_{x \in D} |f(x)|$ for $f \in B(D)$, i.e., the space of bounded functions on D.

Corollary 3.1. *Let μ be a finite measure and $H \subset B(D)$ be a normed space of functions on D, and $\ell_1, \dots, \ell_N$ be bounded linear functionals on $B(D)$. Then, for each $V_n \subset H$ with (3.2) for $w_i = \frac{1}{N}$ the corresponding unweighted least squares method from (3.1) satisfies*

$$\Big\| f - A_N(f) \Big\|_{L_2} \;\le\; \frac{c}{\alpha} \inf_{g \in V_n} \|f - g\|_\infty \qquad \text{for all } f \in H$$

with $c \le \sup\{\|f\|_{L_2} + |\ell_i(f)| \colon \; f \in H, \, \|f\|_\infty = 1, \, i = 1, \dots, N\}$.

In particular, if we are free to choose V_n, then we might want to take one that minimizes the right hand side. For finding corresponding functionals in the case of function evaluations, it has been observed in [6, Theorem 6.2] that for every $V_n \subset C(D)$ of dimension n, there exist $x_1, \dots, x_N$

with $N \asymp n$ with (3.2) for $w_i = 1/N$; see also [21, 78, 112] for earlier results. Applied to the (near-)optimal subspace Corollary 3.1 gives

$$\sup_{f\in F} \left\| f - A_N(f) \right\|_{L_2} \lesssim d_n(F, L_\infty)$$

with $A_N(f) = \operatorname{argmin}_{g\in V_n} \sum_{i=1}^N |f(x_i) - g(x_i)|^2$ with the Kolmogorov width as in (2.2), see [6, Coro. 6.4] and [112, Thm. 1.1]. This is proven using slightly larger iid point sets which are subsampled to obtain point sets of optimal size. This technique will be discussed briefly in Section 5.1. Following the same arguments, similar results can be obtained for more general classes of functionals.

For obtaining optimality of the used method A_N among all linear algorithms, it is of interest to replace the L_∞ on the right hand side by L_2 since $d_n(F, L_2) = a_n(F, L_2)$. This is possible under further (necessary) assumptions on F and ℓ_i, and will be the subject of the following sections.

We end this section with the proof of Proposition 3.1, which follows very closely the lines of Section 3 in [73].

Proof of Proposition 3.1. Let $V_n = \operatorname{span}\{b_1, \dots, b_n\} \subset H$ for some orthonormal system in L_2. Then, the algorithm from (3.1) can be written as

$$A_N(f) = \sum_{k=1}^n \big(G^+ N(f)\big)_k \, b_k,$$

where $N\colon F \to \mathbb{R}^N$ with $N(f) := \big(\sqrt{w_i}\,\ell_i(f)\big)_{i\le N}$ is the (weighted) *information mapping* and $G^+ \in \mathbb{R}^{n\times N}$ is the Moore-Penrose inverse of the matrix

$$G := \big(\sqrt{w_i}\,\ell_i(b_k)\big)_{i\le N,\, k\le n} \in \mathbb{R}^{N\times n},$$

whenever G has full rank.

This is fulfilled, because (3.2) is equivalent to $s_n(G) \ge \alpha$, where $s_n(G)$ denotes the n-th singular value of G. In particular, we have $A_N(g) = g$ for every $g \in V_n$ and $\|G^+\|_{2\to 2} = s_n(G)^{-1} \le \frac{1}{\alpha}$. Therefore,

$$\begin{aligned} \|f - A_N(f)\|_{L_2} &\le \|f - g\|_{L_2} + \|g - A_N(f)\|_{L_2} = \|f - g\|_{L_2} + \|A_N(f - g)\|_{L_2} \\ &= \|f - g\|_{L_2} + \|G^+ N(f - g)\|_{\ell_2^n} \\ &\le \|f - g\|_{L_2} + \frac{1}{\alpha} \|N(f - g)\|_{\ell_2^N}, \end{aligned}$$

which proves the claim. □

3.2 Hilbert spaces and (random) matrices

The very general result in Proposition 3.1 has useful implications for Hilbert spaces. In the following, we consider a separable Hilbert space H which is continuously embedded into $L_2 = L_2(\mu)$, where μ is a measure on some set. We shall assume that the norm of H is given by

$$\|f\|_H^2 := \sum_{k=1}^\infty |\langle f, \sigma_k b_k\rangle_H|^2 = \sum_{k=1}^\infty \frac{|\langle f, b_k\rangle_{L_2}|^2}{\sigma_k^2}, \tag{3.3}$$

where $\{b_1, b_2, \dots\}$ is an orthogonal basis of H that is orthonormal in $L_2(\mu)$, and $\sigma_1 \ge \sigma_2 \ge \cdots > 0$. Note that by the spectral theorem such a basis exists and $\sigma_n \to 0$ holds, if the embedding $H \hookrightarrow L_2(\mu)$ is compact, see e.g. [95, Section 4.2.3]. Further, it holds that $d_k(B_H, L_2) = a_k(B_H, L_2) = c_k(B_H, L_2) = \sigma_{k+1}$ for all $k \in \mathbb{N}$.

If we choose $V_n = \operatorname{span}\{b_1, \dots, b_n\}$ as the *optimal subspace* of H, see e.g. [102, Thm. IV.2.2.], we obtain the following consequence of Proposition 3.1, see e.g. [54].

Proposition 3.2. *Let $H \hookrightarrow L_2$ be a separable Hilbert space with norm as in (3.3). Moreover, for $N \geq n$, let $\ell_1, \dots, \ell_N$ be continuous linear functionals on H, P_n be the orthogonal projection onto $V_n := \operatorname{span}\{b_1, \dots, b_n\}$, and assume that*

$$\left(\sum_{i=1}^{N} w_i \, \ell_i(b_k) \, \overline{\ell_i(b_j)} \right)_{k,j=1}^{n} \geq \alpha^2 \, I_n \tag{3.4}$$

and

$$\left(\sum_{i=1}^{N} w_i \, \ell_i(\sigma_k b_k) \, \overline{\ell_i(\sigma_j b_j)} \right)_{k,j=n+1}^{\infty} \leq \beta^2 \, I, \tag{3.5}$$

for some $\alpha > 0$, $\beta \geq 0$ and weights $w_1, \dots, w_N \geq 0$. Then, for all $f \in H$, the algorithm from (3.1) with the corresponding V_n, ℓ_i and w_i satisfies

$$\left\| f - A_N(f) \right\|_{L_2} \leq \left(\sigma_{n+1} + \frac{\beta}{\alpha} \right) \|f - P_n f\|_H. \tag{3.6}$$

For the worst-case error over the unit ball B_H of H we obtain from (3.6) that

$$e(A_N, B_H, L_2) \leq \sigma_{n+1} + \frac{\beta}{\alpha}. \tag{3.7}$$

Since we always have $e(A_N, B_H, L_2) \geq \sigma_{N+1}$, this leads to an optimal bound (up to constants) provided that $\frac{\beta}{\alpha} \lesssim \sigma_{n+1}$ with $N \asymp n$. Later, we will see that this can be obtained whp for certain classes of random information. But first we consider a general result that works for many more classes of information if we allow a *logarithmic oversampling*.

Let us also mention that it is well known that in the setting of Proposition 3.2 *optimal information* N_n^* is given by $\ell_i(\cdot) = \langle b_i, \cdot \rangle$, $i = 1, \dots, n$. In this case, we can choose $N = n$ and equal weights $w_i = 1$ to have $\alpha = 1$ and $\beta = 0$. Then, in fact, $A_N = P_n$ and equality holds in (3.7), i.e., $e(N_n^*, B_H, L_2) = \sigma_{n+1}$. This recovers the best possible bound for approximation in Hilbert spaces, see e.g. [95, Thm. 4.11].

The above results show that the least squares algorithm from (3.1) satifies "good" error bounds for all $f \in H$ at once, if weights and functionals can be found that satisfy (3.4) and (3.5) with large α and small β. For later use we restate these conditions in the form

$$s_n\Big(\big(\sqrt{w_i} \, \ell_i(b_j) \big)_{1 \leq i \leq N, 1 \leq j \leq n} \Big) \geq \alpha \tag{3.8}$$

and

$$s_1\Big(\big(\sqrt{w_i} \, \ell_i(\sigma_j b_j) \big)_{1 \leq i \leq N, j > n} \Big) \leq \beta, \tag{3.9}$$

where $s_k(A)$ denotes the k-th singular value of a matrix A. In the following, we will see that the existence of good (iid) information can be guaranteed for rather general classes of information Λ and Hilbert spaces H. This is based on results from *random matrix theory*, in particular the concentration result for sums of rank-one matrices in Lemma 3.3 below, which we shall apply to get (3.4) and (3.5) whp. In this way we obtain the "local version" of the main result from [72], but in a slightly more general form, see also [115].

Theorem 3.2. *Let $H \hookrightarrow L_2$ be a Hilbert space with norm as in (3.3) and $\sum_k \sigma_k^2 < \infty$. Moreover, let $\Lambda \subset H'$ be a class of information, and ν be a (possibly infinite) measure on Λ with*

$$\int_\Lambda \ell(f) \cdot \ell(g) \, \mathrm{d}\nu(\ell) \; = \; \langle f, g \rangle_{L_2} \quad \text{for all } f, g \in H. \tag{3.10}$$

Further, let $\ell_1, \dots, \ell_N \in \Lambda$ be iid random functionals distributed with ν-density

$$\rho_n(\ell) \;:=\; \frac{1}{2}\left(\frac{1}{n}\sum_{k\le n}|\ell(b_k)|^2 + \frac{\sum_{k>n}\sigma_k^2\,|\ell(b_k)|^2}{\sum_{k>n}\sigma_k^2}\right), \quad \ell\in\Lambda. \tag{3.11}$$

Then, with probability (at least) $1-N^{-c}$, the algorithm A_N from (3.1) with $V_n = \operatorname{span}\{b_1,\dots,b_n\}$, $N \gtrsim_c n\log n$, functionals ℓ_i and weights $w_i = \rho_n(\ell_i)^{-1}$ satisfies

$$\Big\|f - A_N(f)\Big\|_{L_2} \;\le\; 3\max\left\{\sigma_{n+1},\,\sqrt{\frac{1}{n}\sum_{k>n}\sigma_k^2}\right\}\,\big\|f-P_nf\big\|_H \tag{3.12}$$

for all $f\in H$, where P_n is the orthogonal projection onto V_n. In particular,

$$e_N^{iid}(B_H, L_2, \rho_{32n}\cdot \mathrm{d}\nu) \;\le\; \sqrt{\frac{1}{n}\sum_{k>n}\sigma_k^2}$$

with the same probability if $N \gtrsim_c n\log n$.

The expression $\frac{1}{n}\sum_{k\le n}|\ell(b_k)|^2$ in the density (3.11) is also called (the inverse of) the *Christoffel function*, at least for $\ell(f)=\delta_x(f)$ being given by function evaluations, see e.g. [21] and the references therein.

Another way to write the density is

$$\rho_n(\ell) \;\sim\; \sup_{g\in V_n}\frac{|\ell(g)|}{\|g\|_{L_2}} \;+\; \tau_n\cdot\sup_{g\in H\cap V_n^\perp}\frac{|\ell(g)|}{\|g\|_H}$$

with $\tau_n := (\frac{1}{n}\sum_{k>n}\sigma_k^2)^{-1}$.

To illustrate the effect of the sampling density ρ_n, which also appears inversely in the weights in the algorithm A_N, assume that the $\ell(b_k)$ can attain arbitrarily large and small values. Roughly speaking, while functionals with large values of $\rho_n(\ell)$ are likely to be sampled, their contribution in the algorithm is "damped" and vice versa for functionals for which the density small. It is not entirely clear why these competing effects are beneficial.

We recently learned that a density similar to (3.11) has been used in [4] in the context of *random feature expansions*. Hence, this density appears useful also in the context of constructing "good" random subspaces for approximation, in contrast to our emphasis on finding "good" information.

Let us also add that sampling according to ρ_n might be non-trivial. However, in many cases, like for Gaussian information or for function evaluations for certain Sobolev spaces one can sample instead from the underlying measure ν.

We end this section by providing the proofs of Proposition 3.2 and Theorem 3.2.

Proof of Proposition 3.2. Since

$$\left(\sum_{i=1}^N w_i\,\ell_i(b_k)\,\overline{\ell_i(b_j)}\right)_{k,j=1}^n \;\ge\; \alpha^2\, I_n$$

is equivalent to (3.2), i.e.,

$$\inf_{g\in V_n}\frac{\sqrt{\sum_{i=1}^N w_i\,|\ell_i(g)|^2}}{\|g\|_{L_2}} \;\ge\; \alpha,$$

we obtain from Proposition 3.1 with $g := P_nf$ that

$$\begin{aligned}\Big\|f - A_N(f)\Big\|_{L_2} \;&\le\; \big\|f-P_nf\big\|_{L_2} \;+\; \frac{1}{\alpha}\sqrt{\sum_{i=1}^N w_i\;|\ell_i(f-P_nf)|^2}\\ &\le\; \sigma_{n+1}\,\|f-P_nf\|_H + \frac{1}{\alpha}\sqrt{\sum_{i=1}^N w_i\;|\ell_i(f-P_nf)|^2}.\end{aligned}$$

Note that for every $f \in H$ we have $f - P_n f \in H \cap V_n^\perp$ and, by continuity, $\ell_i(f - P_n f) = \sum_{k>n} \langle f, \sigma_k b_k \rangle_H \, \ell_i(\sigma_k b_k)$, $i = 1, \dots, N$. From

$$\left(\sum_{i=1}^N w_i \, \ell_i(\sigma_k b_k) \, \overline{\ell_i(\sigma_j b_j)} \right)_{k,j=n+1}^{\infty} \; \le \; \beta^2 \, I,$$

which is equivalent to

$$\sup_{g \in H \cap V_n^\perp} \frac{\sqrt{\sum_{i=1}^N w_i \, |\ell_i(g)|^2}}{\|g\|_H} \; \le \; \beta, \tag{3.13}$$

we obtain

$$\sqrt{\sum_{i=1}^N w_i \, |\ell_i(f - P_n f)|^2} \; \le \; \beta \, \|f - P_n f\|_H$$

which completes the proof. □

The proof of Theorem 3.2 is based on the following concentration inequality for infinite matrices, which was essentially proved in [86, Theorem 2.1], see also [1, 98, 114]. Here we use a tailored reformulation of this result from [115, Prop. 1].

Lemma 3.3. *Let $N \ge 3$, $R > 0$ and $y_1, \dots, y_N$ be i.i.d. random sequences from $\ell_2(\mathbb{N})$ satisfying $\|y_i\|_2^2 \le R^2$ almost surely and $\|E\|_{2\to 2} \le 1$, where $E = \mathbb{E}(y_i y_i^*)$.*

Then

$$\mathbb{P}\left(\left\| \frac{1}{N} \sum_{i=1}^N y_i y_i^* - E \right\| \; > \; \frac{1}{2} \right) \; \le \; \frac{4}{N^c},$$

whenever

$$\frac{N}{\ln(N)} \; \ge \; 64(2+c) \, R^2,$$

e.g., if $N \ge 2^9(1+c)\ln(e+c)R^2 \ln(eR^2)$ for $R \ge 2$.

(Numbers in the last line have been checked numerically.)

Proof of Theorem 3.2. We follow the lines of [72] (see also [29]).

In order to apply Proposition 3.2, it remains to show that (3.4) and (3.5) hold for some $\alpha, \beta > 0$ with sufficiently high probability. Define the random vectors

$$(y_i)_k = \begin{cases} \sqrt{w_i} \, \ell_i(b_k), & \text{if } 1 \le k \le n, \\ \dfrac{\sqrt{w_i} \, \ell_i(\sigma_k \, b_k)}{\gamma_n}, & \text{if } k > n, \end{cases}$$

where

$$\gamma_n \; := \; \max\left\{ \sigma_{n+1}, \, \sqrt{\frac{1}{n} \sum_{k>n} \sigma_k^2} \right\} > 0.$$

By the choice of the weights $w_i = \rho_n(\ell_i)^{-1}$, we obtain

$$\|y_i\|_2^2 \; = \; w_i \left(\sum_{k \le n} |\ell_i(b_k)|^2 + \gamma_n^{-2} \sum_{k>n} \sigma_k^2 \, |\ell_i(b_k)|^2 \right) \; \le \; 2n.$$

Further,

$$(y_i y_i^*)_{k,j} \; = \; \begin{cases} w_i \, \ell_i(b_k) \, \overline{\ell_i(b_j)}, & \text{if } 1 \le k, j \le n, \\ \dfrac{w_i \, \ell_i(\sigma_k \, b_k) \, \overline{\ell_i(\sigma_j \, b_j)}}{\gamma_n^2}, & \text{if } k > n, \end{cases}$$

and using (3.10) we obtain

$$\mathbb{E} y_i y_i^* \;=\; \begin{pmatrix} I_n & 0 \\ 0 & C \end{pmatrix},$$

where C is an infinite diagonal matrix with spectral norm at most one, and the expectation is with respect to $\rho_n \cdot \mathrm{d}\nu$.

Lemma 3.3 yields

$$\left(\sum_{i=1}^{N} w_i \, \ell_i(b_k) \, \overline{\ell_i(b_j)} \right)_{k,j=1}^{n} \;\geq\; N I_n - \frac{N}{2} I_n \;=\; \frac{N}{2} I_n$$

and

$$\left(\sum_{i=1}^{N} w_i \, \ell_i(\sigma_k b_k) \, \overline{\ell_i(\sigma_j b_j)} \right)_{k,j=n+1}^{\infty} \;\leq\; N\gamma_n^2 (C \,+\, \frac{1}{2} I) \;\leq\; \frac{3N}{2} \gamma_n^2 \, I,$$

simultaneously with probability $1 - 4N^{-c}$, if $N \geq C' n \log n$ for some constant $C' > 0$ that only depends on c. Hence, we have that the algorithm A_N from Proposition 3.2 satisfies, with the same probability, that

$$\Big\| f - A_N(f) \Big\|_{L_2} \;\leq\; (\sigma_{n+1} + 3\gamma_n) \, \big\| f - P_n f \big\|_H$$

for all $f \in H$. Using that

$$\gamma_n^2 = \max \left\{ \sigma_{n+1}^2, \, \frac{1}{n} \sum_{k>n} \sigma_k^2 \right\} \;\leq\; \frac{2}{n} \sum_{k > n/2} \sigma_k^2,$$

one obtains the bound

$$\Big\| f - A_N(f) \Big\|_{L_2} \;\leq\; \sqrt{\frac{32}{n} \sum_{k \geq \lceil n/2 \rceil} \sigma_k^2} \, \big\| f - P_n f \big\|_H.$$

Replacing n by $32n$ and taking the supremum over $f \in B_H$ implies the uniform bound over B_H. □

3.3 General classes of functions

The purpose of this and the following section is to transfer the above results for L_2-approximation in Hilbert spaces to more general situations. In fact, it will turn out that the "local" result from the last section (referring to the expression $\|f - P_n f\|_H$ on the right hand side of (3.12)) can be used directly to prove bounds for more general classes of functions.

The main idea is to construct a Hilbert space H containing F, and then apply Theorem 3.2 to H. In the context of random information, this approach was used already in [73]. Another construction of such an H, which is also the one we employ here, has been found recently in [68].

Unfortunately, we need that the information functionals ℓ_i are continuous on this Hilbert space, which seems difficult to formalize in a nice form. In fact, as the last proof showed, we need that $\ell(f) = \sum_{k \in \mathbb{N}} \langle f, b_k \rangle_{L_2} \, \ell(b_k)$ holds for all $f \in F$ (or a dense subset) for almost all $\ell \in \Lambda$ with respect to the chosen measure on Λ. Here, we work under the following assumption.

Assumption A. We consider the following setting.

(A.1) F is a separable metric space of functions on a set D.

(A.2) μ is a measure on D such that F is continuously embedded into $L_2 := L_2(\mu)$. Λ is a class of functionals on $\mathcal{L}_2(\mu)$, the μ-square integrable functions on D.

(A.3) ν is a measure on Λ such that ν-almost every (ν-a.e.) $\ell \in \Lambda$ is continuous on F and for every $f \in L_2$ the map $\ell \mapsto \ell(f)$ belongs to $L_2(\nu)$, in particular is well-defined, and

$$\int_\Lambda \ell(f) \cdot \ell(g) \, \mathrm{d}\nu(\ell) \;=\; \langle f, g \rangle_{L_2}, \qquad f, g \in L_2.$$

Assumption (A.3) seems to be quite restrictive, but it fits the examples we have in mind, like function evaluations or Gaussian information, see Section 4. However, other classes of information, like derivative values or *local averages* of the input, see e.g. [13], do not satisfy this assumption, meaning that we cannot find a measure ν such that (A.3) holds. Note that (A.3) readily implies that, for all $f, g \in L_2$, we have $\ell(f+g) = \ell(f) + \ell(g)$ for ν-a.e. $\ell \in \Lambda$, see the proof of Lemma 3.5.

Remark 3.1. *Note that in IBC, or numerical analysis in general, a common assumption is that the information functionals must be defined solely on the class F, or a "surrounding" normed space $H \supset F$. Our analysis, however, requires that we can extend a.e. $\ell \in \Lambda$ also to more general subspaces of L_2. This led us to Assumption (A.3), which may be weakened to equality up to constants independent of $f, g \in L_2$. It would be interesting to find for given F necessary conditions on Λ such that (A.3) holds.*

Under the above assumption we obtain the following general statement.

Theorem 3.4. *There are constants $b, C \in \mathbb{N}$ such that for any F, D, μ, Λ, ν that satisfy Assumption A and all $n \in \mathbb{N}$ and $N \geq Cn\log(n+1)$, we have*

$$e_N^{iid}(F, L_2, \rho_{bn} \cdot \mathrm{d}\nu) \;\leq\; \frac{1}{\sqrt{n}} \sum_{k \geq n} \frac{d_k(F, L_2)}{\sqrt{k}}$$

with probability $1 - N^{-42}$ where the density ρ_n is given by (3.11), with suitable $\{b_k\}$ and $\sigma_k := k^{-1/2} d_{\lfloor k/8 \rfloor}(F, L_2)$. The bound is achieved by the corresponding algorithm A_N from Theorem 3.2.

Recall that $d_k(F, L_2) = a_k(F, L_2)$ if $F \subset L_2$, and that $d_n(F, L_2) < \varepsilon$ is equivalent to the existence of an n-dimensional $V_n \subset F$ with $\sup_{f \in F} \inf_{g \in V_n} \|f - g\|_{L_2} < \varepsilon$.

The following technical lemma is a composition from Section 6.1 of [29] and the proof of [68, Prop. 11], extended to general information.

Lemma 3.5. *Let Assumption A be fulfilled and assume $d_k(F, L_2) > 0, k \in \mathbb{N}$, with*

$$\sum_{k \in \mathbb{N}} \frac{d_k(F, L_2)}{\sqrt{k}} \;<\; \infty.$$

Then, there is an ordered orthonormal system $(b_k)_{k \in \mathbb{N}}$ in L_2 such that $F \subset H \subset L_2$, where the Hilbert space H is the completion of $\operatorname{span}\{b_k\}_{k \in \mathbb{N}}$ with respect to the norm

$$\|f\|_H^2 \;:=\; \sum_{k=1}^{\infty} \frac{\sqrt{k}\,|\,\langle f, b_k\rangle_{L_2}\,|^2}{d_{\lfloor k/8 \rfloor}(F, L_2)}$$

and, for each $n \in \mathbb{N}$, we have

$$\sup_{f \in F} \|f - P_n f\|_H \;\leq\; 4 \sqrt{\sum_{k > \lfloor n/8 \rfloor} \frac{d_k(F, L_2)}{\sqrt{k}}}, \tag{3.14}$$

where P_n is the orthogonal projection onto $V_n := \operatorname{span}\{b_1, \dots, b_n\}$.

Moreover, for every countable set $F_0 \subset F$, there is some $\Lambda_0 \subset \Lambda$ with $\nu(\Lambda \setminus \Lambda_0) = 0$ such that

$$\ell(f) \;=\; \sum_{k=1}^{\infty} \langle f, b_k\rangle_{L_2}\, \ell(b_k) \tag{3.15}$$

for all $f \in F_0$ and $\ell \in \Lambda_0$.

(The modifications in case of $d_k(F, L_2) = 0$ for $k \geq k_0$ are straightforward.)

Before we prove this lemma, let us see how it implies Theorem 3.4.

Proof of Theorem 3.4. We want to apply Theorem 3.2 to the Hilbert space H from Lemma 3.5. In order to extend Λ to linear and continuous functionals on H, let us fix a countable dense subset

$F_0 \subset F$, and corresponding Λ_0 as in Lemma 3.5, and (formally) define, for each $\ell \in \Lambda_0$, the functionals $\overline{\ell}(f)\colon H \to \mathbb{R}$ by $\overline{\ell}(f) = \sum_{k\in\mathbb{N}} \langle f, b_k\rangle_{L_2}\, \ell(b_k)$.

In order to show the boundedness/continuity of $\overline{\ell}\colon H \to \mathbb{R}$, note that

$$|\overline{\ell}(f)| \;\le\; \sum_{k\in\mathbb{N}} |\,\langle f, b_k\rangle_{L_2}\, \ell(b_k)| \;\le\; \|f\|_H \sqrt{\sum_{k\in\mathbb{N}} \sigma_k^2\, \ell(b_k)^2}$$

with $\sigma_k^2 := k^{-1/2} d_{\lfloor k/8\rfloor}(F, L_2)$. Using

$$\int_\Lambda \sum_{k=1}^{n} \sigma_k^2 \ell(b_k)^2 \,\mathrm{d}\nu(\ell) \;\le\; \sum_{k\in\mathbb{N}} \sigma_k^2 \;<\; \infty,$$

for $n \in \mathbb{N}$, we obtain with the monotone convergence theorem that there is $\Lambda' \subset \Lambda$ with $\nu(\Lambda\backslash\Lambda') = 0$ such that $\sum_{k\in\mathbb{N}} \sigma_k^2 \ell(b_k)^2 < \infty$ for all $\ell \in \Lambda'$.

Thus, for every $\ell \in \Lambda_1 := \Lambda_0 \cap \Lambda'$, the functional $\overline{\ell}\colon H \to \mathbb{R}$ is linear and continuous on H, and we set $\overline{\Lambda} := \{\overline{\ell}\colon \ell \in \Lambda_1\} \subset H'$.

We can now apply Theorem 3.2 to H and $\overline{\Lambda}$, where we equip $\overline{\Lambda}$ in the natural way with the same measure as Λ. Equation (3.10) then follows from Assumption (A.3). We obtain, with probability $1 - N^{-c}$, that

$$\Big\|f - A'_N(f)\Big\|_{L_2} \;\le\; \sqrt{\frac{1}{n}\sum_{k\ge n} \frac{d_k(F, L_2)}{\sqrt{k}}} \cdot \|f - P_{32n} f\|_H$$

for all $f \in H$, if $N \gtrsim_c n\log(n+1)$ and $A'_N\colon H \to V_{32n}$ is the algorithm from Theorem 3.2 with iid functionals $\overline{\ell}_i \in \overline{\Lambda}$ according to $\rho_{32n}\,\mathrm{d}\nu$.

Now, for the algorithm A_N as described in Theorem 3.4, note that $\ell_i \in \Lambda_1$ almost surely and, by Lemma 3.5, $\overline{\ell} = \ell$ on F_0 for all $\ell \in \Lambda_1$. This implies that "$A_N(f) = A'_N(f)$ for all $f \in F_0$" with probability one. We obtain that

$$\sup_{f\in F_0} \Big\|f - A_N(f)\Big\|_{L_2} \;\le\; \frac{4}{\sqrt{n}} \sum_{k\ge n} \frac{d_k(F, L_2)}{\sqrt{k}}$$

with probability $1 - N^{-c}$ if $N \gtrsim_c n\log(n+1)$, where we also used (3.14). Since F_0 is dense in F and both $\mathrm{id}\colon F \to L_2$ and $A_n\colon F \to L_2$ are continuous ν-a.e., the same is true for F. □

Proof of Lemma 3.5. From [73, Lemma 3], using that $d_n(F, L_2) = a_n(F, L_2)$ for all $F \subset L_2$, we find an ordered orthonormal system $(b_k)_{k\in\mathbb{N}}$ in L_2 such that

$$\sup_{f\in F} \|f - P_n f\|_{L_2} \;\le\; 2\, d_{\lfloor n/4\rfloor}(F, L_2)$$

for all $n \in \mathbb{N}$. In particular, every $f \in F$ expands into $f = \sum_{k\in\mathbb{N}} \langle f, b_k\rangle_{L_2}\, b_k$ in L_2.

Let H be the Hilbert space in the statement of Lemma 3.5, i.e., we complete $\operatorname{span}\{b_k\}_{k\in\mathbb{N}}$ with respect to the norm

$$\|f\|_H^2 := \sum_{k\in\mathbb{N}} \frac{|\,\langle f, b_k\rangle_{L_2}\,|^2}{\sigma_k^2} \qquad \text{with } \sigma_k^2 := \frac{d_{\lfloor k/8\rfloor}(F, L_2)}{\sqrt{k}}, \quad k \in \mathbb{N}.$$

Then, writing $d_k := d_k(F, L_2)$ and $\alpha_n := \sup_{f\in F} \|f - P_n f\|_2$, we have $\alpha_k \le 2d_{\lfloor k/4\rfloor}$ and obtain for

$n \in \mathbb{N}$ and $f \in F$ that

$$\begin{aligned}\|f - P_n f\|_H^2 &= \sum_{k>n} \frac{\sqrt{k}\,\left|\langle f, b_k\rangle_{L_2}\right|^2}{d_{\lfloor k/8\rfloor}} \le \sum_{\ell=0}^{\infty} \frac{\sqrt{2^{\ell+1}n}}{d_{\lfloor 2^{\ell-2}n\rfloor}} \sum_{k=2^\ell n+1}^{2^{\ell+1}n} \left|\langle f, b_k\rangle_{L_2}\right|^2 \\ &\le \sum_{\ell=0}^{\infty} \sqrt{2^{\ell+1}n} \cdot \frac{\alpha_{2^\ell n}^2}{d_{\lfloor 2^{\ell-2}n\rfloor}} \le 2\sum_{\ell=0}^{\infty} \sqrt{2^{\ell+1}n} \cdot d_{\lfloor 2^{\ell-2}n\rfloor} \\ &\le 2\sum_{\ell=0}^{\infty} \sqrt{2^{\ell+1}n} \cdot \frac{1}{2^{\ell-3}n} \sum_{k=\lfloor 2^{\ell-3}n\rfloor+1}^{2^{\ell-2}n} d_k \\ &\le 8\sqrt{2}\sum_{\ell=0}^{\infty} \sum_{k=\lfloor 2^{\ell-3}n\rfloor+1}^{2^{\ell-2}n} \frac{d_k}{\sqrt{k}} \le 12 \sum_{k>\lfloor n/8\rfloor} \frac{d_k}{\sqrt{k}}.\end{aligned}$$

In particular, $\lim_{n\to\infty} P_n f = f \in H$.

For (3.15), note that, by Assumption (A.3), the mapping $I_\Lambda \colon L_2(D,\mu) \to L_2(\Lambda,\nu)$, $I_\Lambda f(\ell) := \ell(f)$ is an isometry, and hence linear and injective on $L_2(D,\mu)$. Thus, $\{I_\Lambda b_k\}_{k\in\mathbb{N}}$ is an orthonormal system in $L_2(\Lambda,\nu)$, and for every $f \in L_2(D,\mu)$ we have $f^\Lambda = \sum_{k\in\mathbb{N}} \langle f, b_k\rangle_{L_2}\, b_k^\Lambda$ in $L_2(\Lambda,\nu)$, where we write $f^\Lambda := I_\Lambda f$. Every $f \in F$ satisfies

$$\sum_{k\ge 1} k\,|\,\langle f, b_k\rangle_{L_2}\,|^2 = \sum_{n\ge 0}\sum_{k>n} |\,\langle f, b_k\rangle_{L_2}\,|^2 = \sum_{n\ge 0} \|f - P_n f\|_{L_2}^2 < \infty,$$

and hence, by the Rademacher-Menchov theorem, see e.g. [85, 105], we obtain

$$\ell(f) = f^\Lambda(\ell) = \sum_{k\in\mathbb{N}} \langle f, b_k\rangle_{L_2}\, b_k^\Lambda(\ell) = \sum_{k\in\mathbb{N}} \langle f, b_k\rangle_{L_2}\, \ell(b_k) \tag{3.16}$$

for ν-almost all $\ell \in \Lambda$. Since F_0 is countable, the almost everywhere convergence holds simultaneously for all $f \in F_0$, i.e., there is $\Lambda_0 \subset \Lambda$ with $\nu(\Lambda \setminus \Lambda_0) = 0$ and (3.16) for all $f \in F_0$ and $\ell \in \Lambda_0$. In particular, $\ell(b_k)$ is well-defined for $\ell \in \Lambda_0$ and $k \in \mathbb{N}$. □

3.4 Approximation in general norms

We will now present the techniques from [68] to transfer the L_2-error bounds to more general seminorms. However, the results come with additional restrictions on the used approximation spaces.

Assumption B. We consider the following setting.

(B.1) G is a seminormed space which contains F, and $G \cap L_2$ is complete w.r.t. the natural seminorm $\|\cdot\|_* := \|\cdot\|_G + \|\cdot\|_{L_2}$. If two functions from G are equal μ-almost everywhere, then their seminorm in G is the same.

(B.2) $(V_n)_{n=1}^\infty$ is a sequence of subspaces of $G \cap L_2$, respectively of dimension n.

Assumption (B.1) is satisfied, e.g., if F is a compact subset of $C(D)$, where D is a compact domain, μ is a finite measure on D, and $G = L_q(\mu)$ for $1 \le q \le \infty$. Note that for probability measures, and $q \le 2$, the above presented upper bounds for L_2-approximation directly transfer to L_q-approximation via $\|\cdot\|_{L_q} \le \|\cdot\|_{L_2}$, and in also to the integration problem. See [68] for the discussion of more general norms, in particular $G = C(D)$.

We introduce the quantities

$$B_n := B(V_n, G) := \sup_{f\in V_n,\, f\neq 0} \frac{\|f\|_G}{\|f\|_2} \tag{3.17}$$

and

$$\alpha_n := \alpha(V_n, F) := \sup_{f\in F} \|f - P_{V_n} f\|_2 = \sup_{f\in F} \inf_{g\in V_n} \|f - g\|_2.$$

Note that B_n corresponds to the (inverse) of the Christoffel function if $G = L_\infty$. Further, for optimal subspaces the quantity $B(V_n, G)$ is equal to

$$\inf_{\dim(V_n)=n} \sup_{f\in V_n,\, f\neq 0} \frac{\|f\|_G}{\|f\|_2} = b_{n-1}(\mathrm{id}\colon G \to L_2)^{-1},$$

where b_n denotes the *Bernstein n-width*, see [102, Ch. II].

The following lemma is easy to prove. We refer to [68, Lemma 10].

Lemma 3.6. *Let Assumption B hold. For any $n \in \mathbb{N}$, any mapping $A\colon F \to V_n$ and all $f \in F$, we have*

$$\|f - Af\|_G \le 2 \sum_{k>\lfloor n/4\rfloor} \frac{\alpha_k\, B_{4k}}{k} + B_n \cdot \|f - Af\|_2.$$

This bound can now be used in combination with the algorithm and sampling strategy from above. We only state a special case, and refer to [68] for the general result, the proof and a discussion of some cases where this leads to sharp results (up to the logarithmic oversampling).

Theorem 3.7. *Let Assumptions A and B hold and let α, β, γ and δ be real parameters with $\alpha > \max\{\beta, 1/2\}$. If*

$$B(V_n, G) \lesssim n^\beta (\log n)^\delta \quad \text{and} \quad \alpha(V_n, F) \lesssim n^{-\alpha}(\log n)^\gamma$$

then

$$e_N^{iid}(F, L_2, \rho_n \cdot \mathrm{d}\nu) \lesssim n^{-\alpha+\beta}(\log n)^{\gamma+\delta}$$

with probability $1 - N^{-\alpha}$ if $N \gtrsim n \log n$, where ρ_n is as in Theorem 3.4.

Note that this includes the (sharp) corollary of Theorem 3.4 for $G = L_2$: In this case $\beta = \delta = 0$, and with the choice of (V_n) such that $\alpha(V_n, F) \le 2d_{\lfloor n/4\rfloor}(F, L_2)$, we obtain that $d_n(F, L_2) \asymp n^{-\alpha}(\log n)^\gamma$ with $\alpha > 1/2$ implies

$$e_{n\log n}^{iid}(F, G, \rho_n \cdot \mathrm{d}\nu) \lesssim d_n(F, L_2) = a_n(F, L_2)$$

with probability $1 - n^{-\alpha}$ whenever Assumption A holds.

3.5 Sharpness for L_2-approximation

Let us note that Theorem 3.4 may be applied to the setting in Section 3.2 and in particular Theorem 3.2, where $d_k(B_H, L_2) = \sigma_{k+1}$. A qualitative comparison gives bounds on $e_N^{iid}(B_H, L_2, \Lambda)$ whp, where $N \gtrsim n \log n$, which are of order

$$\frac{1}{\sqrt{n}} \sum_{k>n} \frac{\sigma_k}{\sqrt{k}} \quad \text{vs.} \quad \frac{1}{\sqrt{n}} \sqrt{\sum_{k>n} \sigma_k^2},$$

where apart from the common prefactor the first expression is the Lorentz $\ell_{2,1}$-norm and the second the ℓ_2-norm of the tail. In general, we have $\ell_{2,1} \subset \ell_{2,2} = \ell_2$, see e.g. [101, Prop. 2.1.10]. That is, if the second sum converges, then so does the first. Let us give some more details on the convergence.

In Section 4, we shall be interested in sequences (σ_n) with $\sigma_n \asymp n^{-\alpha}(\log n)^{-\beta}$ for $\alpha, \beta > 0$, which is the case for n-widths of Sobolev embeddings, see e.g. [102, Ch. VII] or [111, Ch 5]. In this case,

$$\frac{1}{\sqrt{n}} \sum_{k>n} \frac{\sigma_k}{\sqrt{k}} \asymp_{\alpha,\beta} \begin{cases} \sigma_n & \text{if } \alpha > 1/2, \\ \sigma_n \log n & \text{if } \alpha = 1/2 \text{ and } \beta > 1, \end{cases}$$

and

$$\frac{1}{\sqrt{n}}\sqrt{\sum_{k>n}\sigma_k^2} \asymp_{\alpha,\beta} \begin{cases} \sigma_n & \text{if } \alpha > 1/2, \\ \sigma_n\sqrt{\log n} & \text{if } \alpha = 1/2 \text{ and } \beta > 1/2, \end{cases}$$

and in the remaining cases the sum does not converge. So, the second bound is slightly smaller for poly-log decay.

We obtain that $n \log n$ iid measurements are asymptotically optimal whp, i.e., $e^{iid}_{n\log n}(B_H, L_2, \Lambda) \asymp \sigma_n$, provided that the singular numbers decay at a rate $n^{-\alpha}$ with $\alpha > 1/2$. At $\alpha = 1/2$ we lose logarithmic factors. But also these bounds are sometimes sharp up to the oversampling, see [29,74]. In certain cases, the oversampling can be removed, see Section 4. In others, it is necessary, as we shall see in the following.

For this, assume that H is as in Theorem 3.2 and $\Lambda := \{\langle\cdot, h\rangle : h \in \mathcal{B}\}$, where $\mathcal{B}$ is some orthonormal basis of L_2. Note that Assumption (A.3) is fulfilled if ν is the counting measure on Λ. Theorem 3.2 now implies that $n \log n$ random coefficients w.r.t. $\mathcal{B}$ are as powerful as the optimal information for Hilbert spaces.

Now, if $\mathcal{B} = \{b_1, b_2, \dots\}$ is the optimal basis as in (3.3), then the density with respect to the counting measure ν on $\{\langle\cdot, b_k\rangle : k \in \mathbb{N}\}$ is given by

$$\rho_n(\langle\cdot, b_\ell\rangle) = \frac{1}{2}\begin{cases} \frac{1}{n} & \text{if } \ell \le n, \\ \frac{\sigma_\ell^2}{\sum_{k>n}\sigma_k^2} & \text{if } \ell > n. \end{cases}$$

Further, by virtue of the coupon collectors theorem, the asymptotics $n \log n$ cannot be improved. More precisely, one needs on average $n \log n$ samples drawn according to ρ_n (or any other distribution) to guarantee that the first n Fourier coefficients are evaluated. For completeness, we give a formal statement below. A similar effect occurs if we approximate Sobolev functions using function samples, see Section 4.2. This shows that in general Theorem 3.2 and thus Theorem 3.4 cannot be improved.

Proposition 3.3. *Assume that H is as in Theorem 3.2 and $\Lambda := \{\langle\cdot, h\rangle : h \in \mathcal{B}\}$, where $\mathcal{B}$ is some orthonormal basis of L_2. Let ν be any probability measure on Λ. If $N/n\log n \to 0$, then for any $\alpha > 0$ we have*

$$e^{iid}_N(B_H, L_2, \nu) \ge \sigma_{\lfloor \alpha n \rfloor}$$

whp for all large enough n.

If the sequence (σ_n) decays fast enough, i.e., if $\lim_{K\to\infty}\sup_{n\in\mathbb{N}}\frac{\sigma_{Kn}}{\sigma_n} = 0$ such as for polynomial decay, then under the conditions in the proposition we deduce that for any $C > 0$

$$e^{iid}_N(B_H, L_2, \nu) \ge C\sigma_n$$

whp for all n large enough. Thus, for sampling Fourier coefficients, $N \ll n \log n$ iid functionals are whp worse by an arbitrary factor than n optimal functionals.

This shows that for some classes of information that clearly contain optimal information, we need a logarithmic oversampling. This might not come as a surprise. However, we will see in Section 4 that this is not true in general, depending on the classes of information and inputs. This leads to the following question.

Open Problem 3.1. *What are conditions on Λ and H such that the conclusion of Proposition 3.3 holds?*

This should be compared to Theorems 4.5 and 4.4, and Open Problem 4.2 below.

Proof of Proposition 3.3. We identify Λ with $\mathbb{N}$ and thus N iid functionals $\ell_1, \dots, \ell_N$ in Λ sampled according to ν correspond to N numbers sampled randomly from $\mathbb{N}$.

Since $N/n\log n \to 0$, for any n large enough we have $N \le \frac{1}{2}n\log n$. By the coupon collector's problem, the probability that we miss one of the numbers in $\{1, \dots, n\}$, say i, is at least $1 - e^{-\sqrt{n}}$.

For this, combine the classical limit law in [33] with the fact that equidistribution on $\{1,\dots,n\}$ stochastically needs the least amount of coupons, see [12, p.52]. In this case, the i-th Fourier coefficient is not measured and any algorithm A_N lacking this information has error at least $\sigma_i \geq \sigma_n$ (since $\pm\sigma_i b_i \in B_H$, it has to return zero for these functions).

Thus, $e_N^{iid}(B_H, L_2, \nu) \geq \sigma_n$ in this case. By replacing n with $\lfloor \alpha n \rfloor$ for $\alpha > 0$ we get that with probability at least $1 - e^{-\sqrt{\alpha n}}$ for all large enough n we have

$$e_N^{iid}(B_H, L_2, \nu) \geq \sigma_{\lfloor \alpha n \rfloor}.$$

□

4 Applications and improvements

The results above show that $n \log n$ pieces of iid information are often as valuable as optimal information, and we have even seen that this cannot be improved in general. However, there are cases where the logarithmic oversampling factor can be removed and iid information is asymptotically optimal. We report on two instances. Namely, L_q-approximation in (isotropic) Sobolev spaces W_p^s with $p > q$ if iid uniform samples are used, and L_2-approximation if the iid information is *Gaussian*, in both a linear and a nonlinear setting. We still do not completely understand what makes these settings special in this respect, and add some open problems related to them.

We start with a more detailed explanation of the general case of standard information, i.e., function evaluations. In particular, we discuss a natural limitation of this class of information.

4.1 Random function evaluations

Approximation of (regular or smooth) functions using function evaluations was the main motivation and application of the results introduced in Section 3. Clearly, for $\Lambda = \Lambda^{\mathrm{std}}$, Assumption (A.3) is obvious by identifying ν with μ (from $L_2(\mu)$) by $\nu(\{\delta_x \colon x \in M\}) = \mu(M)$ for μ-measurable $M \subset D$. For applying the results we would like point evaluation to be continuous on H.

Especially Hilbert spaces of this type are of interest and intensively studied:

A Hilbert space H of functions on a set D is called a *reproducing kernel Hilbert space (RKHS) on* D if point evaluation $\delta_x \colon H \to \mathbb{R}$ is a continuous functional for all $x \in D$, i.e., H is a RKHS if and only if $\Lambda^{\mathrm{std}} \subset H'$. The crucial property of RKHS is the existence of a (reproducing) kernel $K \colon D \times D \to \mathbb{R}$ such that $f(x) = \langle f, K(x,\cdot)\rangle_H$ for all $f \in H$ and $x \in D$. If $H \hookrightarrow L_2(\mu)$ is compact, the kernel characterizes H in the sense of (3.3) being equivalent to

$$K(x,y) \;=\; \sum_{k=1}^{\infty} \sigma_k^2 \, b_k(x)\, b_k(y), \qquad x,y \in D, \tag{4.1}$$

where, again, $(\sigma_k) \in \ell_2$ is a non-increasing sequence and $\{b_k\}_{k=1}^{\infty}$ is an orthonormal system in L_2, see e.g. [108, Thm. 3.1]. We refer to [2] for more on RKHS.

Let us note that the results from the last section do not require *every* δ_x to be in H'; almost all would suffice. However, this is not a major restriction, as we can always pass to the (full-measure) subset $D_0 \subset D$, where function evaluation is continuous. This does not change the L_2-error, or the random sampling.

Again, we obtain from Theorem 3.2, see [72], that there is some sampling density ρ_n, such that $n \log n$ random sampling points are enough for a near-optimal bound. In particular, the algorithm

$$A_N(f) \;=\; \operatorname*{argmin}_{g \in V_n} \sum_{i=1}^{N} \frac{|g(x_i) - f(x_i)|^2}{\rho_n(x_i)}$$

for suitable V_n and ρ_n from (3.11) (with $\ell = \delta_x$) satisfies

$$e(A_N, B_H, L_2) \;\le\; \sqrt{\frac{1}{n}\sum_{k>n} c_k(B_H, L_2)^2},$$

with probability $1 - N^{-c}$, whenever $N \gtrsim_c n \log n$, and $x_1, \dots, x_N$ are iid w.r.t. $\rho_n \, \mathrm{d}\mu$.

It has already been observed in [58] that the square-summability of (c_n) is a necessary condition for a general comparison with (g_n). In fact, it is shown in [74] that there exists some $c > 0$ such that

$$g_{\lfloor cn \rfloor}(B_H, L_2) \;\ge\; \sqrt{\frac{1}{n}\sum_{k>n} c_k(B_H, L_2)^2}$$

for some (simple, univariate) Hilbert space H and all n, see (2.4). That is, the upper bound is optimal, up to the logarithmic oversampling factor, and even for optimal function evaluations the square-summability of (c_n) is necessary. In fact, for many important examples, as the Sobolev spaces discussed below, this summability corresponds to the embedding into $C(D)$. It is therefore necessary to work with function evaluations and only a weak restriction. We will see in Section 4.3 that the same restriction appears for Gaussian information.

So, in the case of standard information in a RKHS, iid information is optimal up to the logarithmic oversampling factor. Similar results exist for more general classes, see [29, 73].

Remark 4.1. *($g_n \neq c_n$?). It is easy to find examples where standard information is as powerful as arbitrary linear information. For this consider the (pathological) example of $D = \mathbb{N}$ and a RKHS $H \subset \ell_2$ on $\mathbb{N}$ as in (3.3) with norm $\|f\|_H = \sum_{k=1}^{\infty} f_k^2 \sigma_k^{-2}$, for $f = (f_1, f_2, \dots)$, i.e., the ONB $\{b_k\}_{k=1}^{\infty}$ is given by the canonical basis of ℓ_2. Clearly, function evaluation is the same as computing coefficients w.r.t. the (optimal) basis (b_k), and so $g_n(B_H, L_2) = c_n(B_H, L_2)$ in this case.*

However, when we turn to approximation in other norms, then it seems that, so far, no general "for all H" comparison has been observed, and additional conditions appear; often involving the quantity $B(V_n, G)$ from (3.17). Let us only discuss the case of *uniform approximation* $G = L_\infty$, and refer to [68] for generalizations. In this case, we have

$$B(V_n) \;:=\; B(V_n, L_\infty) \;=\; \sup_{f \in V_n,\, f \neq 0} \frac{\|f\|_\infty}{\|f\|_2} \;=\; \Big\| \sum_{k=1}^{n} |b_k|^2 \Big\|_\infty^{1/2},$$

where $\{b_1, \dots, b_n\}$ is an arbitrary L_2-orthonormal basis of V_n.

As Theorem 3.7 shows, bounds on $B(V_n)$, together with good L_2-approximation properties of V_n, leads to a corresponding upper bound on the error of A_N from above. However, for finite μ, we have

$$B(V_n, L_\infty) \;\ge\; \Big(\frac{1}{\mu(D)} \sum_{k=1}^{n} \|b_k\|_2^2\Big)^{1/2} = \sqrt{\frac{n}{\mu(D)}},$$

which shows that we lose at least a factor of $\sqrt{n}$ compared to the L_2-error.

For RKHS on finite measure spaces with $B(V_n) \asymp \sqrt{n}$ for the optimal subspaces V_n we have the following result which is essentially Theorem 6 in [68] and was obtained independently in [39].

Theorem 4.1. *There are absolute constants $b, c \in \mathbb{N}$ such that the following holds. Let μ be a finite measure on a set D and $H \hookrightarrow L_2$ be a reproducing kernel Hilbert space with kernel as in (4.1) with $\sigma_{2n} \gtrsim \sigma_n$ and*

$$\sup_n \Big\| \frac{1}{n} \sum_{k=1}^{n} |b_k|^2 \Big\|_\infty \;<\; \infty.$$

Then, $H \hookrightarrow L_\infty$ and the unweighted least squares method

$$A_N^u(f) \;:=\; \operatorname*{argmin}_{g \in V_n} \sum_{i=1}^{N} |g(x_i) - f(x_i)|^2 \tag{4.2}$$

with $V_n = \operatorname{span}\{b_1, \dots, b_n\}$ *and* $x_1, \dots, x_N \overset{\text{iid}}{\sim} \mu$ *satisfies*

$$e(A_N^u, B_H, L_\infty) \lesssim c_n(B_H, L_\infty) = a_n(B_H, L_\infty)$$

with probability $1 - N^{-c}$, *whenever* $N \gtrsim_c n \log n$.

Note that no decay condition on (c_n) and no sampling density depending on a basis is needed - the unweighted least squares algorithm is *universal*. We will see in the following section that for L_q-approximation with $q < 2$ the logarithmic oversampling can be removed in the case of Sobolev spaces.

The assumptions of Theorem 4.1 hold for example if the basis $\{b_k\}$ is bounded which is the case for the trigonometric system or the Chebychev system or (Haar) wavelets, if μ is their corresponding orthogonality measure. This includes many interesting RKHSs such as certain Sobolev spaces. See [68] for several examples.

Still, it is not clear how far this result can be extended.

Open Problem 4.1. *Find necessary conditions on* H *such that the conclusion of Theorem 4.1 holds. Moreover, find a variant for more general* $F \hookrightarrow L_\infty$.

Note that Theorem 4.1 is only implicitly contained in [39,68] as both papers work directly with the optimal *subsampled* algorithm from [29]. (We discuss this shortly in Section 5.1). However, since the proof is based on a variant of Lemma 3.6, see Section 3.2 of [68], it is apparent that one may also work with the algorithm A_N from Theorem 3.2.

Another ingredient in the above theorem is the next result that allows for removing the weights from algorithm and sampling. We state it for future reference.

Proposition 4.4. *Let* H, μ *and* $\Lambda = \Lambda^{\text{std}}$ *be as in Theorem 4.1. Then the conclusion of Theorem 3.2 and consequently Theorems 3.4 and 3.7 continue to hold for the sampling density* $\varrho_n \equiv \mu(D)^{-1}$, $n \in \mathbb{N}$.

Note that constant weights can be replaced by 1 in (3.1) and thus the algorithm is an unweighted least squares method as in (4.2).

Proof of Proposition 4.4. We consider Theorem 3.2 for H, μ and Λ^{std} as in the statement of the proposition. Then $\Lambda^{\text{std}} \subset H'$ and (3.10) hold. The sampling density enters the proof of Theorem 3.2 in the estimate $\|y_i\|_2^2 \le 2n$. Using the density $\rho \equiv \frac{1}{\mu(D)}$, and therefore $w_i = \mu(D)$, instead, we see that $\|y_i\|_2^2 \le 2C\mu(D)\, n$ a.s. is implied by $|\rho_n(x)| \le C$ for μ-almost all $x \in D$ and all n. We can therefore apply Lemma 3.3 with the corresponding R.

Regarding the first summand in (3.11), we have for μ-almost all $x \in D$ that

$$\frac{1}{n}\sum_{k=1}^{n} |b_k(x)|^2 \le \Big\| \frac{1}{n}\sum_{k=1}^{n} |b_k|^2 \Big\|_{L_\infty(\mu)} = \frac{1}{n} B(V_n, L_\infty(\mu))^2 \lesssim 1.$$

To investigate the second summand in (3.11), let $n \in \mathbb{N}$ and pick $\ell \in \mathbb{N}$ such that $2^\ell \le n \le 2^{\ell+1}$. Then, for μ-almost all $x \in D$,

$$\sum_{k \ge n} \sigma_k^2 |b_k(x)|^2 \le \sum_{k \ge 2^\ell} \sigma_k^2 |b_k(x)|^2 \le \sum_{i=\ell}^{\infty} \sigma_{2^i}^2 \sum_{k=2^i}^{2^{i+1}-1} |b_k(x)|^2$$

$$\lesssim \sum_{i=\ell}^{\infty} \sigma_{2^i}^2 2^{i+1} \lesssim \sum_{i=\ell}^{\infty} \sum_{k=2^{i-1}}^{2^i-1} \sigma_k^2 \lesssim \sum_{k \ge n/4} \sigma_k^2.$$

It remains to use $\sum_{k \ge n/4} \sigma_k^2 \lesssim \sum_{k \ge n} \sigma_k^2$ which follows from assuming $\sigma_{2n} \gtrsim \sigma_n$. □

4.2 Sharp results for Sobolev spaces

In this section we take a closer look at L_q-approximation in isotropic Sobolev spaces for which we have a characterization of the quality of (random) samples due to [70, 71] which implies asymptotic optimality of n or $n \log n$ iid measurements depending on the parameters involved. There are also generalizations to similarly structured isotropic function spaces such as Holder, Triebel-Lizorkin or Besov spaces.

On a domain $D \subset \mathbb{R}^d$ (i.e., an open and nonempty set), equipped with the Lebesgue measure, the Sobolev space of smoothness $s \in \mathbb{N}$ and integrability $1 \leq p \leq \infty$ is given by

$$W_p^s(D) := \Big\{ f \in L_p(D) \colon \|f\|_{W_p^s(D)} := \Big(\sum_{|\alpha| \leq s} \|D^\alpha f\|_{L_p(D)}^p \Big)^{1/p} < \infty \Big\},$$

where the sum is over all multi-indices $\alpha \in \mathbb{N}_0^d$ with $|\alpha| = \alpha_1 + \ldots + \alpha_d \leq s$ and $D^\alpha f = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} f$ denotes a weak partial derivative of order $|\alpha|$. In the following, we denote by $B_p^s(D)$ the unit ball of $W_p^s(D)$.

Sobolev functions from $W_p^s(D)$ do have well-defined function values if the embedding $W_p^s(D) \hookrightarrow C_b(D)$ into the bounded continuous functions holds, i.e., if

$$s > d/p \quad \text{if } 1 < p \leq \infty \quad \text{or} \quad s \geq d \quad \text{if } p = 1, \tag{4.3}$$

and $D \subset \mathbb{R}^d$ is a bounded Lipschitz domain, and if $s > d/p$ the embedding is compact, see, e.g., [84, Sec. 1.4.5]. Then the sampling numbers, i.e., the minimal worst-case errors based on function values, are known to satisfy

$$g_n(B_p^s, L_q) \asymp n^{-s/d + (1/p - 1/q)_+}, \tag{4.4}$$

where $(x)_+ = \max\{0, x\}$.

These asymptotics are classical for special domains like the cube and have been obtained with linear algorithms, see e.g. [92] and the references therein.

Let us apply the general results from above in the special case of $p = 2$. Then W_2^s is a Hilbert space and since the embedding $W_2^s \hookrightarrow C_b$ is compact and D is bounded, also $W_2^s \hookrightarrow L_2$ is compact. By the spectral theorem, W_2^s is of the form (3.3) with

$$\sigma_{n+1} = c_n(B_2^s, L_2) \asymp n^{-s/d},$$

see e.g. [92, Thm. 26]. Since $s > d/2$, Theorem 3.2 gives that whp $N \asymp n \log n$ iid points sampled according to the density ρ_{32n} are as powerful as n optimal points. In order to apply Proposition 4.4 and in particular to conclude the same result with constant sampling density, it is sufficient to have $B(V_n) \lesssim \sqrt{n}$ for (almost) optimal subspaces.

This is for example the case if the domain is a compact Riemannian manifold M of dimension d, where the eigensystem of the Laplace-Beltrami operator provides such subspaces. Combining Proposition 4.4 with Corollary 31 in [68] gives that the unweighted least squares method A_N^u from (4.2) using $N \asymp n \log n$ iid random points sampled according to the normalized uniform measure μ_M on M achieves

$$e_{n \log n}^{iid}(B_2^s, L_q, \mu_M) \lesssim g_n(B_2^s, L_q)$$

whp for all $1 \leq q \leq \infty$. We will see in the following that the logarithmic oversampling is necessary if $q \geq 2$ and can be removed otherwise if we use a particular "localized" least squares method.

For simplicity, in the remainder of this section, the domain D will be a bounded convex domain (and in particular Lipschitz) and we refer to [71] for (almost) analogous results on manifolds. We will suppress D in the notation.

Given a point set $\mathcal{P}_n = \{x_1, \dots, x_n\} \subset D$ we identify it with the corresponding evaluations. The following characterization of the n-th minimal error of iid information, that is, iid points sampled according to the uniform measure μ_D on D, is taken from [67, Thm. 2] (see [70, Cor. 2] for the original result) and conjectured already in [53], where the case $d = s = 1$ has been obtained.

Theorem 4.2. *Let $1 \le p, q \le \infty$ and $s \in \mathbb{N}$ as in (4.3). Then,*

$$\mathbb{E}\, e_n^{iid}\big(B_p^s, L_q, \mu_D\big) \asymp \begin{cases} g_{n/\log n}\big(B_p^s, L_q\big) & \text{if } q \ge p, \\ g_n\big(B_p^s, L_q\big) & \text{if } q < p. \end{cases}$$

Let us note that this result also holds with high probability. The following questions are obvious:

Open Problem 4.2. *Do the bounds of Theorem 4.2 also hold for general classes F? In particular, under which conditions on q and $F \subset L_q(\mu)$ do we have asymptotic optimality of iid function evaluations, i.e., $\mathbb{E}\, e_n^{iid}(F, L_q, \mu) \asymp g_n(F, L_q)$? Moreover, is logarithmic oversampling necessary, i.e., do we have $\mathbb{E}\, e_n^{iid}(B_H, L_2, \mu) \asymp g_{n/\log n}(B_H, L_2)$ for any RKHS H, and $\mathbb{E}\, e_n^{iid}(F, L_\infty, \mu) \gtrsim g_{n/\log n}(F, L_\infty)$ for more general $F \subset L_\infty$?*

Remark 4.2. *At this point, it seems worthwhile noting that in [92] the authors also concluded that, if one restricts to linear methods, linear information can be asymptotically better than standard information if and only if $p < 2 < q$.*

The algorithm achieving the upper bound in Theorem 4.2 is linear and is based on the *moving least squares* method applied to cones adapted to local density of the sampled point set. For more details we refer to [67] and [119, Ch. 4].

In order to describe the algorithm, we introduce a geometric regularity condition on the domain. We say that a set $D \subset \mathbb{R}^d$ satisfies an *interior cone condition* with radius $r > 0$ and angle $\theta \in (0, \pi/2)$ if, for all $x \in D$, there is a direction $\xi(x) \in \mathbb{S}^{d-1}$ such that the cone

$$C(x, \xi(x), r, \theta) := \Big\{ x + \lambda y \colon y \in \mathbb{S}^{d-1}, \langle y, \xi(x) \rangle \ge \cos\theta, \lambda \in [0, r] \Big\}$$

with apex x is contained in D. Convex sets satisfy this condition and also bounded Lipschitz domains, see [70] for proofs and references. Additionally, we can and do assume that $\theta \le \pi/5$ and that ξ depends continuously on x for almost all $x \in D$.

In the following we will describe the algorithm for a fixed point set $P = \{x_1, \dots, x_n\} \subset D$. We can later insert any realization of a random point set. We shall assume that P is sufficiently dense in D.

Given $f \in C_b(D)$ and $x \in D$ we approximate $f(x)$ by

$$A_P f(x) := \operatorname*{argmin}_{v \in V_m} \sum_{y \in P \cap K_P(x)} w(x, y)\, |f(y) - v(y)|^2,$$

where V_m is the space of real polynomials of degree at most $m = \lceil s \rceil$, $K_P(x) := C(x, \xi(x), r_P(x), \theta)$ and the radius $r_P(x) > 0$ is minimal such that there are sufficiently many points in $K_P(x)$ to reconstruct all polynomials in V_m. Further, the weight function takes the form $w(x, y) = \Phi(x - y)$ where Φ is supported in $B_2^n(0, \delta)$ and positive on $B_2^n(0, \delta/2)$, where δ depends on $P \cap K_P(x)$.

Thus for evaluating the approximant $A_P f$ at $x \in D$ we solve a weighted least squares problem depending on the density of points around x; hence, the terminology "moving least squares".

Open Problem 4.3. *Is there an unweighted least squares algorithm such that the bounds of Theorem 4.2 hold?*

By Proposition 4.4 we know that this is the case for $q \ge 2 = p$, at least for manifolds.

Note that Theorem 4.2 holds in fact on all domains satisfying the interior cone condition, see [67, Thm. 2]. However, for bounded convex domains there is a convenient characterization of the radius of information which explains why random points are sometimes optimal and sometimes not.

To this end, introduce the covering radius $h_{\mathcal{P}_n,D} := \sup_{x\in D} \operatorname{dist}(x,\mathcal{P}_n)$ which is the supremum of the distance function

$$\operatorname{dist}(\cdot,\mathcal{P}_n)\colon \mathbb{R}^d \to [0,\infty), \qquad \operatorname{dist}(x,\mathcal{P}_n) := \min_{y\in\mathcal{P}_n} \|x-y\|_2$$

to the n-point sampling set $\mathcal{P}_n \subset D$. Although commonly used, the covering radius is insufficient to characterize the power of information as the following result taken from [70, Thm. 0.1] shows.

Proposition 4.5. *Let $1 \le p,q \le \infty$ and $s \in \mathbb{N}$ as in (4.3). For any point set $\mathcal{P}_n \subset D$, we have*

$$r\big(\mathcal{P}_n, B_p^s, L_q\big) \asymp \begin{cases} \big\| \operatorname{dist}(\cdot,\mathcal{P}_n)\big\|_{L_\infty(D)}^{s-d(1/p-1/q)} & \text{if } q \ge p, \\ \big\| \operatorname{dist}(\cdot,\mathcal{P}_n)\big\|_{L_\gamma(D)}^{s} & \text{if } q < p, \end{cases}$$

where $\gamma = s(1/q-1/p)^{-1}$ and the implicit constants are independent of $\mathcal{P}_n$.

Thus, the quality of a point set is asymptotically determined by the radius of the largest hole amidst the points if $q \ge p$ and by an average of the distance to the point set if $q < p$. Partial results have been obtained in [53,88,92,99,109].

Theorem 4.5 is a tool to analyze the asymptotic optimality of arbitrary (sequences of) point sets and, in particular, random or typical ones. To compare, the optimal behaviour of the L_γ-norm of the distance function is

$$\inf_{\#\mathcal{P}_n \le n} \|\operatorname{dist}(\cdot,\mathcal{P}_n)\|_{L_\gamma(D)} \asymp n^{-1/d} \qquad \text{for every} \quad 0 < \gamma \le \infty.$$

By Theorem 4.5, point sets attaining this rate yield the upper bound in (4.4).

For uniform random points on a bounded convex domain, that is, iid points distributed according to μ_D, it is known that the average hole size is on average of optimal order $n^{-1/d}$, see e.g. [43, Theorem 9.2], whereas the largest hole is on average of size $n^{-1/d}(\log n)^{1/d}$ and thus slightly larger than optimal, essentially due to the coupon collectors' problem, see e.g. [103, Corollary 2.3]. This provides an explanation for Theorem 4.2.

It is natural to end this section with the following questions:

Open Problem 4.4. *What can be used in place of $\operatorname{dist}(\cdot,\mathcal{P}_n)$ to derive a "geometric" characterization of good point sets for other classes F, such as unit balls in anisotropic Sobolev spaces?*

We now turn to random information on Λ^{all}, which does not have a limitation in the sense of optimal information.

4.3 Gaussian information

A geometric problem that was actually the starting point of the renewed interest in random information in the IBC community, see [53,54], is the classical problem of recovering vectors from a symmetric convex body (a compactum with nonempty interior) $K \subset \mathbb{R}^m$ in the norm of ℓ_2^m by using n linear measurements $\ell_1,\dots,\ell_n$ with n much smaller than m. This fits the above setting by choosing $F = K$, which is the unit ball of a normed space $(\mathbb{R}^m, \|\cdot\|_K)$, $G = \ell_2^m$ and $\Lambda = \{\langle\cdot, y\rangle : y \in \mathbb{R}^m\}$. (Note that we can consider vectors as functions on a discrete set.)

In this case, the radius of information $r((\ell_i)_{i=1}^n, K, \ell_2^m)$, see (1.3), has a geometric interpretation since it is equal up to a factor of 2 to the radius

$$\operatorname{rad}(K \cap E_n) := \sup_{x\in K\cap E_n} \|x\|_2,$$

of K intersected with the subspace $E_n = \{x \in \mathbb{R}^m\colon \ell_1(x) = \cdots = \ell_n(x) = 0\}$, see e.g. [95, Lem. 4.3]. If the measurements are linearly independent, then E_n is of codimension $n < m$, i.e., of dimension $m-n$, and the smallest possible radius corresponds to the Gelfand numbers/width $c_n(K,\ell_2^m)$, see (2.3).

It is natural to ask:

How large is a "typical" intersection, if we choose the subspace E_n uniformly at random?

A canonical choice of a uniform distribution is the normalized Haar measure on the set of all subspaces of codimension $n < m$, i.e. on the Grassmannian manifold $\mathcal{G}_{n,m}$. It turns out that if N_n is a Gaussian matrix with independent standard Gaussian entries, that is if we choose the standard Gaussian measure γ_m on $\mathbb{R}^m$, then $E_n^{\mathrm{ran}} = \ker N_n$ is distributed according to this measure, and that is why we focus on this *Gaussian information*. The radius of the intersection of a convex body $K \subset \mathbb{R}^m$ with such a random subspace therefore satisfies

$$\mathrm{rad}(K \cap E_n^{\mathrm{ran}}) \asymp e_n^{iid}(K, \ell_2^m, \gamma_m), \tag{4.5}$$

where the implicit constant is independent of any realization.

The above is a classical and well-studied question, which was tackled by many authors, especially for n of the order m, i.e., intersections of large codimension. See e.g. the classical results [40,41,81], or the recent findings [42,80] which were obtained in the context of asymptotic geometric analysis, see [3] for additional references.

In the following we apply the above results to ellipsoids of the form

$$\mathcal{E}_\sigma = \left\{ (x_1, x_2, \dots) \in \ell_2 \colon \sum_{k\in\mathbb{N}} \sigma_k^{-2} x_k^2 \le 1 \right\}$$

with square-summable semi-axes $\sigma_1 \ge \sigma_2 \ge \dots \ge 0$, i.e., $\sum \sigma_j^2 < \infty$. In this case $\mathcal{E}_\sigma$ is the unit ball of a separable Hilbert space $H \hookrightarrow \ell_2$. Note that this includes the finite-dimensional case, where we set $\sigma_k = 0$ for $k > m$ and demand then that $x_k = 0$.

Let $\{e_1, e_2, \dots\} \subset \ell_2$ be the standard basis and $g_1, g_2, \dots$ be iid standard Gaussian random variables. The sequence $(g_1, g_2, \dots)$ is distributed on $\mathbb{R}^{\mathbb{N}}$ according to the countable product of standard Gaussian measure. We define a Gaussian random functional by $f \mapsto \ell(f) = \sum_{j=1}^\infty \langle f, e_j\rangle_2 g_j$ for $f \in H$ which almost surely absolutely converges and is therefore in H'. Then the restriction γ of the distribution of ℓ to H' is a centered Gaussian measure on H' and (3.10) holds. Gaussian information is universal in the sense that it is invariant under change of basis, i.e., we have $\ell(f) = \sum_{j=1}^\infty \langle f, u_j\rangle_2 \, g_j$ in distribution for any other orthonormal basis $\{u_j\}$ and $f \in H$. We refer to [10] for details.

In order to apply Theorem 3.2, note that for every $f, h \in H$,

$$\begin{aligned}\int_{H'} \langle f, g\rangle_2 \langle h, g\rangle_2 \,\mathrm{d}\gamma(g) &= \mathbb{E}\Big(\sum_{j=1}^\infty \langle f, e_j\rangle_2 g_j\Big)\Big(\sum_{j=1}^\infty \langle h, e_j\rangle_2 g_j\Big) \\ &= \sum_{j=1}^\infty \langle f, e_j\rangle_2 \langle h, e_j\rangle_2 = \langle f, h\rangle_2\end{aligned}$$

and we get a bound on $e_N^{iid}(H, L_2, \rho_{32n} \cdot \mathrm{d}\nu)$ with $N \gtrsim n \log n$ whp. For Theorem 3.2, we choose the random functionals $\ell(\cdot) = \sum_{j=1}^\infty \langle \cdot, e_j\rangle_2 v_j$ by choosing the coefficients $(v_k)_{k\in\mathbb{N}}$ w.r.t. the density $\rho_n \,\mathrm{d}\gamma$, where

$$\rho_n(\ell) = \frac{1}{2}\Big(\frac{1}{n}\sum_{k\le n} |v_k|^2 + \sum_{k>n} \beta_k |v_k|^2\Big) \quad \text{with } \beta_k = \frac{\sigma_k^2}{\sum_{k>n}\sigma_k^2}.$$

This density concentrates around 1 (with respect to γ) as $n \to \infty$, see e.g. [77, Lem. 1]. In order to apply Lemma 3.3 we need an almost sure bound and so we cannot use density equal to one which would correspond to the geometric setting. In the following, we will see that we can choose a constant density and remove the logarithmic oversampling.

Recall that for $\ell_1, \dots, \ell_N \overset{\mathrm{iid}}{\sim} \gamma$ and $f \in H$ we have $\ell_i(f) = \sum_{j=1}^\infty \langle f, e_j\rangle_2 g_{ij}$ for iid standard Gaussians g_{ij} and thus $\ell_i(e_k) = g_{ik}$, i.e., the matrices in (3.8) and (3.9) will be structured Gaussian.

Following [54], we use matrix concentration bounds for these random matrices and in the following lemma combine the lower bound on the n-th singular value from [25, Thm. II.13] and the upper bound on the first from [5, Cor. 3.11] for simplicity in the special case $N = 2n$.

Lemma 4.3. *Let $n \in \mathbb{N}$ and $N = 2n$. Consider $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ with $\sigma \in \ell_2$. There exists $c > 0$ such that with probability $1 - 2e^{-cn}$ we have*

$$s_n\Big(\big(N^{-1/2}\, g_{ij}\big)_{1\leq i\leq N, 1\leq j\leq n}\Big) \;\geq\; \frac{1}{2}$$

and

$$s_1\Big(\big(N^{-1/2}\, \sigma_j g_{ij}\big)_{1\leq i\leq N, j>n}\Big) \;\leq\; 2\sqrt{\frac{1}{n}\sum_{j>n}\sigma_j^2} + 2\sigma_{n+1}.$$

Combining Lemma 4.3 with Proposition 3.2, see also (3.8) and (3.9), implies that with probability $1 - 2e^{-cn}$ the least squares algorithm A_N in (3.1) using $N = 2n$ Gaussian random functionals $\ell_1, \dots, \ell_N$ and weights $1/N$ has error on $f \in H$ bounded by

$$\|f - A_N(f)\|_{\ell_2} \;\leq\; 5\Big(\sigma_{n+1} + \sqrt{\frac{1}{n}\sum_{j>n}\sigma_j^2}\Big)\, \|f - P_n f\|_H. \tag{4.6}$$

Due to $\sigma_{n+1} + \sqrt{\frac{1}{n}\sum_{j>n}\sigma_j^2} \lesssim \sqrt{\frac{1}{n}\sum_{j>n/2}\sigma_j^2}$ we deduce the following result obtained in [54, Thm. 3].

Theorem 4.4. *There are absolute constants $b, c \in \mathbb{N}$ such that, for all $\sigma \in \ell_2$, we have that bn Gaussian measurements satisfy*

$$e_{bn}^{iid}(\mathcal{E}_\sigma, \ell_2, \gamma) \;\leq\; \sqrt{\frac{1}{n}\sum_{j>n}\sigma_j^2}$$

with probability $1 - e^{-cn}$.

In [54] even a lower bound was shown such that for $\sigma_k \asymp n^{-\alpha}(\log n)^{-\beta}$ with $\alpha > 0$ and $\beta \in \mathbb{R}$ the characterization

$$e_n^{iid}(\mathcal{E}_\sigma, \ell_2, \gamma) \;\asymp_{\alpha,\beta}\; \begin{cases} \sigma_1 & \text{if} \quad \alpha < 1/2 \text{ or } \beta \leq \alpha = 1/2, \\ \sigma_n\sqrt{\log n} & \text{if} \quad \beta > \alpha = 1/2, \\ \sigma_n & \text{if} \quad \alpha > 1/2, \end{cases}$$

holds with high probability, see the proof of Corollary 7 in [54]. It turns out that the n-th minimal error of iid Gaussian information is of the same order as the minimal radius if $\sigma \in \ell_2$, while random information seems useless if $\sigma \notin \ell_2$. Note that this is exactly the threshold we have seen for (random) standard information, see Section 4.1.

Thus, in the case of Hilbert spaces we have a complete picture of the power Gaussian information for ℓ_2-approximation, at least for poly-logarithmic decay, and the corresponding algorithms are linear.

In contrast, not much is known about the case of ℓ_p-approximation.

Open Problem 4.5. *Investigate $e_n^{iid}(\mathcal{E}_\sigma, \ell_p, \gamma)$ for $p \neq 2$, or approximation in more general norms.*

In the following, we briefly discuss implications for non-Hilbert spaces. For simplicity, we restrict ourselves to the better known finite-dimensional case.

For the unit ball K of a normed space $(\mathbb{R}^m, \|\cdot\|_K)$, Assumption A is satisfied if μ is the counting measure on $\{1, \dots, m\}$ and ν the standard Gaussian measure on $\mathbb{R}^m$. By Lemma 3.5 we find a suitable Hilbert space in $\mathbb{R}^m$ such that for each $n \in \mathbb{N}$ and $x \in K$ we have

$$\|x - P_n x\|_H \;\lesssim\; \sum_{k=\lfloor n/8\rfloor+1}^{m} \frac{d_k(K, \ell_2^m)}{\sqrt{k}}.$$

If we insert this into (4.6), then we obtain (similarly to the proof of Theorem 3.7) that with some constant oversampling factor $b > 1$ the unweighted least-squares algorithm A_N based on $N = bn$ iid Gaussian random functionals satisfies

$$\|x - A_N(x)\|_2 \lesssim \frac{1}{\sqrt{n}} \sum_{k=n+1}^{m} \frac{d_k(K, \ell_2^m)}{\sqrt{k}},$$

for every $x \in \mathbb{R}^m$ with probability $1 - 2e^{-cn}$, where $c > 0$ is as in Lemma 4.3. Thus, we get the following upper bound.

Corollary 4.5. *Let $K \subset \mathbb{R}^m$ be a convex body. There exist $c, C > 0$ such that*

$$e_n^{iid}(K, \ell_2^m, \gamma_m) \leq \frac{C}{\sqrt{n}} \sum_{k=\lfloor cn \rfloor}^{m} \frac{d_k(K, \ell_2^m)}{\sqrt{k}}$$

holds with probability $1 - 2e^{-cn}$ for all $n \leq m$.

Via (4.5) this upper bound holds for the radius of a typical section. There is a similar bound in terms of Gelfand widths, which correspond to optimal sections and therefore seems better suited as a benchmark.

Proposition 4.6. *([81, Thm. 3.2]). Let $K \subset \mathbb{R}^m$ be a convex body. There exist $c, C > 0$ such that*

$$e_n^{iid}(K, \ell_2^m, \gamma_m) \leq \frac{C}{\sqrt{n}} \sum_{k=\lfloor cn \rfloor}^{m} \frac{c_k(K, \ell_2^m)}{\sqrt{k}}, \tag{4.7}$$

holds with probability $1 - e^{-cn}$ for all $n \leq m$.

The proof of this result relies on a "rounding technique" together with an M^*-estimate, which is also employed for example in [40, 41, 54]. It gives a direct estimate on the radius $\mathrm{rad}(K \cap E_n^{\mathrm{ran}})$ instead of providing an explicit reconstruction algorithm using Gaussian information. Thus, only an abstract nonlinear algorithm can be given which matches the bound (4.7), see Section 5.2.

Using the asymptotics stated in Section 3.5 we can derive that Gaussian information is asymptotically optimal if the Gelfand widths decay a little faster than $n^{-1/2}$, i.e., the bodies are sufficiently "thin" as the dimension increases. It would be interesting whether this threshold of $n^{-1/2}$ is sharp. Again, this is the threshold we have seen for (random) standard information, see Section 4.1.

Open Problem 4.6. *Is there some $K \subset \ell_2$ such that $(c_n(K, \ell_2)) \notin \ell_2$, but still $e_n^{iid}(K, \ell_2, \gamma) \to 0$ a.s.?*

In order to investigate the sharpness of the bound (4.7), in [59] random sections of ℓ_p-ellipsoids have been studied which are images of ℓ_p-balls with $0 < p \leq \infty$ under diagonal operators. In the case $1 < p \leq \infty$ the logarithmic gaps present for poly-log decay with $\alpha = 1/2$ can be narrowed. It would be interesting to do this also for general convex bodies. The proofs behind build on the same techniques used in [54] and [81], and consequently yield a nonlinear algorithm.

Remark 4.3. *(More general linear information). In the finite-dimensional case with a symmetric convex body $K \subset \mathbb{R}^m$ which corresponds to the geometric problem of finding small sections of K we note that the obtained general results not only hold for the Gaussian measure. In fact, to satisfy Assumption (A.3), we can take any measure ν on $\mathbb{R}^m$ which is isotropic in the sense of*

$$\int_{\mathbb{R}^m} \langle x, u \rangle \langle y, u \rangle \, \mathrm{d}\nu(u) = \langle x, y \rangle.$$

If additionally ν has barycenter at the origin, then this corresponds to isotropicity as used in asymptotic geometric analysis, see e.g. [3, Sec. 10.2]. Further note that Lemma 4.3, which works without logarithmic oversampling, also holds for example for Rademacher random variables instead of standard Gaussian ones, see [5] and [104], and thus has implications for other types of information.

5 Further topics

Let us shortly touch upon some topics close to the scope of this survey.

5.1 Subsampling and optimal information

Instead of considering random sampling, and related minimal errors, it is clearly of interest to find relations between the different "benchmarks" of optimal approximation, see Section 2. One particular reason is the study of the *power* of certain classes of information, in which case the "best" information has to be considered.

It is to some extent surprising that the general results of the previous sections already lead to an optimal comparison in some cases. In fact, one can employ a subsampling technique based on the famous solution to the Kadison-Singer problem [82] to reduce a given "good" set of information to an "optimal" subset. We will discuss the essential lemma at the end of this section.

This has been done in [29] in the case of function values, see also [87, 112], and the following theorem is a slight generalization.

Theorem 5.1. *There is a constant $b \in \mathbb{N}$ such that for any F, D, μ, Λ, ν that satisfy Assumption A and all $n \in \mathbb{N}$, we have*

$$e_{bn}(F, L_2, \Lambda) \le \frac{1}{\sqrt{n}} \sum_{k \ge n} \frac{d_k(F, L_2)}{\sqrt{k}}.$$

The bound is achieved by the corresponding algorithm A_N from Theorem 3.2, with the $N \asymp n \log n$ random functionals replaced by a suitable subset of order n. Again, a slight improvement is possible for Hilbert spaces. We omit the details and refer to [29].

Since $d_k(F, L_2) = a_k(F, L_2)$ for all $F \subset L_2$, Theorem 5.1 shows that, whenever the approximation numbers of F decay at a polynomial rate larger $1/2$, then information which satisfies Assumption (A.3) is asymptotically as powerful as arbitrary linear information for L_2-approximation in F, at least if we only allow linear algorithms, see (2.1).

If $F = B_H$ is the unit ball of a Hilbert space, then it is known that $a_n(B_H, L_2) = c_n(B_H, L_2)$, and the last result shows that

$$e_n(B_H, L_2, \Lambda) \asymp c_n(B_H, L_2) \tag{5.1}$$

whenever $c_n(B_H, L_2) \asymp n^{-\alpha}$ for some $\alpha > 1/2$.

That is, the class Λ contains *optimal information*. Recall that this applies, e.g., to Λ^{all}, to coefficients w.r.t. an arbitrary ONB of L_2, or to function evaluations Λ^{std}. The latter case in particularly interesting as it was an open problem for a while. This, and the corresponding open problems from [31, 97], were solved in [29], see also [76, 94, 117] for earlier results on this, and [72, 73, 87, 115] for direct predecessors. It is not clear what makes Λ^{std}, or other information with (A.3), special in this context.

This motivates the following open problem.

Open Problem 5.1. *Find necessary and sufficient conditions on H and Λ, independent of n, such that (5.1) holds true.*

So far, we know that (A.3) together with some decay of (c_n) is sufficient, and that $\sup_{\ell \in \Lambda} |\ell(f)| > 0$ for all $f \neq 0$ is necessary.

Recall that for Λ^{std} relation (5.1) is, in general, not true for Hilbert spaces with $(c_n) \notin \ell_2$, see [58, 74]. In contrast, it is obvious by definition that no condition on the decay of (c_n) is needed to achieve (5.1) for Λ^{all}. It would be interesting to find a class Λ such that we have for some $p^* > 2$, that (5.1) holds for Hilbert spaces H with $\Lambda \subset H'$ and $(c_n) \in \ell_p$ with $p > p^*$, but does not hold for some H with $(c_n) \in \ell_{p^*}$. For example, we do not know the answer if $\Lambda = \Lambda^{\mathrm{coef}}$ consists of coefficients w.r.t. an arbitrary ONB of L_2. The same questions are clearly of large interest for

non-Hilbert spaces, and approximation in other norms. We leave that for future research, and just note that the same subsampling approach was used in [39,68] to obtain the optimal bound

$$g_{bn}^{\mathrm{lin}}(B_H, L_\infty) \asymp c_n(B_H, L_\infty)$$

for all RKHS H that satisfy the conditions of Theorem 4.1.

The final ingredient for the proof of Theorem 5.1 was the following infinite-dimensional version of the subsampling (or *sparsification*) theorem, which allowed for direct application in the above described setting, see Proposition 17 of [29].

Lemma 5.2. [29]. *There are absolute constants $c_1 \leq 43200$, $c_2 \geq 50$, $c_3 \leq 21600$, with the following properties. Let $n, N \in \mathbb{N}$ and $y_1, \dots, y_N$ be vectors from $\ell_2(\mathbb{N}_0)$ satisfying $\|y_i\|_2^2 \leq 2n$ and*

$$\Big\| \frac{1}{N} \sum_{i=1}^{N} y_i y_i^* - \begin{pmatrix} I_n & 0 \\ 0 & \Lambda \end{pmatrix} \Big\|_{2\to 2} \leq \frac{1}{2},$$

with the identity $I_n \in \mathbb{C}^{n\times n}$ and some Hermitian matrix Λ with $\|\Lambda\|_{2\to 2} \leq 1$. Then, there is a subset $J \subset \{1, \dots, N\}$ with $|J| \leq c_1 n$, such that

$$\Big(\frac{1}{n} \sum_{i\in J} y_i y_i^* \Big)_{<n} \geq c_2\, I_n \qquad \text{and} \qquad \frac{1}{n} \sum_{i\in J} y_i y_i^* \leq c_3\, I,$$

where $A_{<n} := (A_{k,l})_{k,l<n}$ and $A \leq B$ denotes the Loewner order.

A short look to the proof of Theorem 3.2 reveals how to apply this lemma to reduce the number of samples from $n \log n$ to n, while preserving the spectral properties of the involved matrices. This result, as its finite-dimensional origin, is fascinating, especially because it does not depend on the initial sample size N. See also [78] for an application to sampling discretization, or [63] for a survey, and [36] for the discretization of continuous frames.

Finally, we remark that Lemma 5.2 is ultimately due to the solution of the Kadison-Singer problem in [82], together with the iterative approach from [89].

5.2 Nonlinear sampling algorithms

Our focus is on linear algorithms but here we want to mention some results regarding nonlinear algorithms using iid information. We refer to the survey [26], or the recent works [9, 17, 19] and references therein, for more information on nonlinear approximation.

In general, if H and G are normed spaces and information is given by a map $N_n \colon H \to \mathbb{R}^n$, then the reconstruction mapping

$$\varphi^*(y) := \operatorname*{argmin}_{g\in G} \sup\{\|g - h\|_G \colon h \in F,\, N_n(h) = y\},$$

if it exists, is optimal, i.e., it attains the infimum in (1.3). In fact, it returns a Chebyshev center of the set $N^{-1}(y) \cap F$ considered as a subset of G. Composed with N_n this gives an optimal nonlinear algorithm A_n^*. As mentioned in Section 2 and seen in Section 4, often linear algorithms using iid information can be asymptotically as good but this is not always the case.

One of the most prominent instances of the success of iid information for nonlinear approximation is the case $F = \ell_1^m$, $G = \ell_2^m$ and F the unit ball of ℓ_1^m which is a special case of the geometric problem mentioned in Section 4.3. This case is related to sparse recovery, see e.g. [30,35], and was resolved already in [62] and [38]. In fact, it was shown that

$$\mathbb{E}\, e_n^{iid}(\ell_1^m, \ell_2^m, \gamma_m) \asymp c_n(\ell_1^m, \ell_2^m) \asymp \min\left\{1, \sqrt{\frac{\log(1 + \frac{m}{n})}{n}}\right\}, \tag{5.2}$$

where the hidden constants are absolute and γ_m denotes the standard Gaussian measure on $\mathbb{R}^m$. Note that the corresponding approximation numbers are much larger and thus nonlinear reconstructions are strictly better, see [100].

We refer to [53] for more details on the proof of (5.2) which is based on ℓ_1-minimization or *basis pursuit* and references to generalizations. Recently, this has been generalized to ℓ_p-ellipsoids with implications for Gelfand numbers of diagonal operators, see [59].

It is remarkable that so far, despite its enormous importance for applications, there is no explicit, deterministic construction of a *near-optimal* N_n attaining the upper bound in (5.2). The same is true for several of the results from Section 3.

Remark 5.1. *Let us note that in the original bound in [62] on $e_n^{iid}(\ell_1, \ell_2, \gamma_m)$ the exponent of $\log(1+\frac{m}{n})$ in (5.2) is $3/2$ instead of $1/2$. It is somehow interesting that, given the optimal bound on $c_n(\ell_1^m, \ell_2^m)$ in (5.2) this can be obtained from the bound in [81] presented in Proposition 4.6.*

In the context of sampling numbers, we want to mention further recent results based on sparse approximation and iid random points.

Using a greedy (and nonlinear) approximation method, as well as iid uniform random points on rather general domains [24] obtained bounds for L_2-approximation in general function classes, thereby improving upon recent results in [61] obtained via *basis pursuit denoising*, another nonlinear reconstruction method. See also [66] for an analysis of this method with emphasis on high-dimensional approximation.

5.3 Randomized algorithms

Randomized algorithms, also known as *Monte Carlo methods*, are a larger class of algorithms which, in contrast to the algorithms discussed so far, are allowed to use different information for each input, and additional random numbers. (Although we studied random information, we considered the deterministic worst-case error for each realization as in (1.3), and hence, do not allow random algorithms in this sense.) That is, a Monte Carlo method M_n is a random variable, that depends in expectation on n pieces of information of the input, and we define the *worst-case (root-mean-square) error*

$$e^{\mathrm{ran}}(M_n, F, G) \;:=\; \sup_{f\in F} \sqrt{\mathbb{E}\Big[\|f - M_n(f)\|_G^2\Big]}$$

as well as the *n-th minimal randomized errors* $e_n^{\mathrm{ran}}(F, G, \Lambda)$ as the infimum over all such methods. We clearly have $e_n^{\mathrm{ran}}(F, G, \Lambda) \le e_n(F, G, \Lambda)$, because every deterministic algorithm can be considered a (constant) random variable. In addition, randomized methods might be quite advantageous and more generally applicable. However, such methods do usually not allow for reliable error guarantees, in the sense that error bounds only hold with certain probability, and that a realization of a randomized algorithm may have small error for some $f \in F$, but not for all at once.

There are even many situations (e.g., if H and G are Hilbert spaces, F is the unit ball of H, and we allow arbitrary linear information) where randomness does not help at all compared to deterministic algorithms. We refer to [48, 90, 95, 96] for more details and general results on randomized approximation.

We will see below that a randomized least squares method can attain the optimal results under much weaker conditions. First, it is intuitively clear, and can be seen from the corresponding proofs, that the "discretization condition" (3.2), or (3.4), is crucial also in this case. But the "stability condition" (3.13), or (3.5), which depends on the class F, can be weakened to

$$\mathbb{P}\left(\sum_{i=1}^N w_i |\ell_i(g)|^2 \;\le\; \beta^2 \,\|g\|_{L_2}^2\right) \;\ge\; 1-\eta \qquad \text{for all } \; g \in L_2, \tag{5.3}$$

for some $\eta > 0$, where w_i and ℓ_i are the random weights and functionals, respectively. This condition is easy to verify for iid ℓ_i if they are sampled, as before, with respect to a density ρ_n', and we set

$w_i = 1/\rho_n'(\ell_i)$.

The following has been obtained essentially in [21].

Proposition 5.7. *Let $V_n \subset L_2$ be an n-dimensional space, and w_i, ℓ_i be such that* (5.3) *and*

$$\mathbb{P}\left(\sum_{i=1}^{N} w_i |\ell_i(g)|^2 \ \geq \ \alpha^2 \, \|g\|_{L_2}^2 \quad \text{for all } \ g \in V_n\right) \ \geq \ 1-\delta \tag{5.4}$$

hold for some $\alpha, \beta, \eta, \delta > 0$. Then, the algorithm A_N from (3.1) *satisfies for each $f \in L_2$ that*

$$\Big\|f - A_N(f)\Big\|_{L_2}^2 \ \leq \ \left(1 + \frac{\beta}{\alpha}\right) d(f, V_n, L_2)^2$$

with probability $1-\eta-\delta$, where $d(f, V_n, G) := \min_{g \in V_n} \big\|f - g\big\|_G$.

We see that, once the conditions are verified, we obtain a near-optimal approximation in arbitrary subspaces of L_2. We do not need to assume that $f \in F$ for some class F with decaying widths. It is clear that a result of this kind cannot be true in a deterministic setting, or with probability one.

To obtain a bound in expectation, i.e., on the error $e^{\mathrm{ran}}(A_N)$, we need to control the error for realizations of A_N for which (5.3) and (5.4) do not hold. This can be done in different ways. In [21], where this result was applied first to $N \asymp n \log n$ iid points, the authors proceeded by considering an error bound in terms of $d(f, V_n, L_\infty)$, or by adding a term $n^{-r}\|f\|_2$ on the right hand side, where N must grow with r, see also [15, 18]. The required sampling density ρ_n' is the first summand of ρ_n in (3.11). In [45], the algorithm was analyzed for iid random points distributed according to this density, conditioned on the event in (5.4). Since (5.3) also holds in expectation, one obtains $\mathbb{E}\|f - A_N(f)\|_{L_2}^2 \lesssim d(f, V_n, L_2)^2$ for $N \asymp n \log n$. Based on a similar subsampling idea as discussed in Section 5.1, this led to the important result from [20] which shows that, in the case of L_2-approximation, linear randomized algorithms based on function values can be optimal among arbitrary linear algorithms. See [64, 118] for earlier results, and [16] for a recent refinement leading to explicit and smaller constants and oversampling.

Theorem 5.3. [20]. *There exist constants $b, C \in \mathbb{N}$ such that the following holds. For any n-dimensional space $V_n \subset L_2(D, \mu)$, there is a random variable X on $\binom{D}{bn}$, i.e., all (bn)-element subsets of D, such that the algorithm A_N from* (3.1) *with $N = bn$, $\{x_1, \ldots, x_{bn}\} = X$ and $w_i := \min_{v \in V_n} \frac{\|v\|_{L_2}^2}{|v(x_i)|^2}$ satisfies*

$$\mathbb{E}\Big\|f - A_N(f)\Big\|_{L_2}^2 \ \leq \ C \min_{g \in V_n} \big\|f - g\big\|_{L_2}^2 \qquad \text{for all } \ f \in L_2.$$

In particular,

$$e_{bn}^{\mathrm{ran}}(F, L_2, \Lambda^{\mathrm{std}}) \ \leq \ C \cdot a_n(F, L_2)$$

for any compact subset $F \subset L_2$.

The random sample used by the above algorithm is not given by iid random points, and it is again not too difficult to see (by using the coupon collector's problem) that such a result cannot be true for iid information in general.

Similar to the results and techniques from Section 4.2, it has been shown in [67] that iid samples are asymptotically optimal in expectation for L_q-approximation of Sobolev functions in W_p^s on very general domains, except for the case $p = q = \infty$. This is an improvement compared to the deterministic error discussed in Section 4.2. Again, we would like to know how this generalizes.

Open Problem 5.2. *Find conditions on F such that there exists a density ρ with $\mathbb{E}\|f - A_N(f)\|_{L_2}^2 \lesssim d(f, V_n, L_2)^2$ for all $f \in F$, where A_N from* (3.1) *is based on $N \asymp n$ iid points from ρ.*

Regarding nonlinear approximation which was discussed in Section 5.2, Gaussian information has proven useful also in the randomized setting for approximation between sequence spaces and

L_q-approximation in Sobolev spaces, see e.g. [47,83]. Recently, in [75], following [49], improvements using adaptivity have been shown, which is in contrast to the deterministic case, where adaption is useless for linear problems, see [95, Thm. 4.4].

5.4 High dimensions and tractability

Many problems in numerical approximation have an associated dimension d, that of the domain, which influences the error. In IBC it is of large interest to study the dependence of the minimal errors on d.

In this context, it can be useful to state results in terms of the *information complexity*

$$n(\varepsilon, F, G, \Lambda) := \min\bigl\{n \in \mathbb{N}\colon\ e_n(F, G, \Lambda) \le \varepsilon\bigr\},$$

which is the minimal number of values of information functionals from Λ needed to achieve an error of at most $\varepsilon > 0$. That is, every algorithm that achieves an error in G of at most $\varepsilon > 0$ for all $f \in F$ needs at least $n(\varepsilon, F, G, \Lambda)$ pieces of information from Λ.

To classify the *difficulty* of a problem in higher dimensions we consider a sequence of function classes F_d on domains D_d (and associated G_d and Λ_d), $d \in \mathbb{N}$. A problem (or a sequence thereof) is then called *polynomially tractable*, if there exist absolute constant $\alpha, \beta, C \ge 0$, such that

$$n(\varepsilon, F_d, G_d, \Lambda_d) \ \le\ C\, d^{\alpha}\, \varepsilon^{-\beta} \qquad \text{for all } \varepsilon > 0 \ \text{ and } \ d \in \mathbb{N},$$

i.e., the needed amount of information depends at most polynomially on d and $1/\varepsilon$. In contrast, a problem is said to suffer from the *curse of dimension*, if there exist absolute constants $\varepsilon_0, d_0, \gamma, C > 0$, such that

$$n(\varepsilon, F_d, G_d, \Lambda_d) \ \ge\ C\,(1+\gamma)^d \qquad \text{for all } \varepsilon \in (0, \varepsilon_0) \ \text{ and } \ d \ge d_0.$$

That is, one needs exponentially many pieces of information to find an approximate solution to the problem. Such a problem is generally assumed to be *intractable*. For a comprehensive account on tractability and the concepts mentioned in the remainder of this section we refer to the books [95–97].

Remark 5.2. *The term* "curse of dimension" *has been introduced by Bellmann [8] in 1957 for the phenomenon that the number of needed samples increases exponentially with the (input) dimension. This is very much inspired by classifications in discrete complexity theory and we use the same concept. Note, however, that this term gained prominence in several other areas and is sometimes used for saying that the order of convergence decreases to zero (e.g., like $e_n \asymp_d n^{1/d}$). The examples below will show that this is not the same, and that the d-dependent "constants" are of significance.*

As a result of the structure of function spaces, in many cases the curse of dimension cannot be avoided. It is largely open, and one of the main concerns of IBC, to identify natural classes of functions where the curse of dimension does not hold, especially for Λ^{std}.

In the context of this survey let us only mention that the spaces $W_p^s(D)$ from Section 4.2 are too large to be tractable. Namely, it has been shown in [55–57], see also [109] for the case $s = 1$, that for any sequence of volume-normalized domains (D_d), like $D_d = [0,1]^d$, already the integration problem for $W_\infty^s(D_d)$ suffers from the curse of dimension. In fact,

$$n(\varepsilon, W_\infty^s(D_d), L_q, \Lambda^{\mathrm{std}}) \ \gtrsim_s\ \left(\frac{d}{\varepsilon}\right)^{d/s}$$

for all $1 \le q \le \infty$, $s \in \mathbb{N}$ and volume-normalized (D_d). Even larger bounds exist for $q = \infty$, see [65], in which case the curse also holds for $s = \infty$, see [93].

As a remedy to the curse of dimension we mention the prominent example of *weighted spaces*, introduced in [106], where different "weights" are assigned to different coordinates, see [95–97] for a discussion. For recent progress in the context of "unweighted" spaces, see [66] where tractability of

L_q-approximation in spaces with bounded sum of absolute values of Fourier coefficients with respect to a suitable basis was obtained.

Note that many of the error bounds presented above come with (explicit) absolute constants and are therefore also suitable for tractabilty studies, see e.g. [39, 69].

There is another prominent problem in IBC, which emphasizes the value of iid information in this context: Distributing points "uniformly" on $[0,1]^d$.

For this, let the (star-)discrepancy of a point set $\mathcal{P}_n$ be given by

$$D(\mathcal{P}_n) = := \sup_{x\in[0,1]^d} \Big| \frac{\#(\mathcal{P}_n \cap [0,x])}{n} - \mathrm{vol}([0,x]) \Big|,$$

where $[0,x] = [0,x_1] \times \cdots \times [0,x_d]$. Optimizing over all n-point sets $\mathcal{P}_n \subset [0,1]^d$ gives the *n-th minimal discrepancy*

$$D(n,d) := \inf_{\mathcal{P}_n} D(\mathcal{P}_n).$$

This is a very important and extensively studied quantity in the field of *irregularity of distribution* [7]. See also [27, 52, 91] for recent treatments.

Via the prominent *Koksma-Hlawka inequality* [60] the discrepancy of $\mathcal{P}_n$ is related to the radius of information for integration in a space of mixed smoothness, see e.g. [27]. To date, the best bounds (for large n) are

$$n^{-1} (\log n)^{(d-1)/2+\eta_d} \lesssim_d D(n,d) \lesssim_d n^{-1} (\log n)^{d-1}, \tag{5.5}$$

with some small $\eta_d > 0$, see [14] for the lower bound and e.g. [27] for the upper bound. Regarding the dependence on d, there exist $c_1, c_2, \varepsilon_0 > 0$ such that

$$c_1 \min\left\{\varepsilon_0, \frac{d}{n}\right\} \le D(n,d) \le c_2 \sqrt{\frac{d}{n}} \qquad \text{for all } d,n \in \mathbb{N}, \tag{5.6}$$

see [50, 51]. We obtain that the number of points needed to achieve a discrepancy less than $\varepsilon > 0$ is (up to constants) between $d\,\varepsilon^{-1}$ and $d\,\varepsilon^{-2}$ and hence, linear in d. The bound from (5.5), which increases exponentially with d, is therefore not enough to conclude a statement on the complexity in high-dimensions.

The upper bound due to [50] is achieved by iid uniform random points and relies on [110] which employs empirical process theory and the concept of *Vapnik-Červonenkis (VC) dimension*, a notion of complexity originating from statistical learning theory. In high dimensions, iid points achieve also the best known bounds for the related notion of *dispersion*, which measures the volume of the largest empty box, see [79, 116].

Note that due to the central limit theorem the rate $n^{-1/2}$ in (5.6) cannot be improved for iid points. Even more, for uniformly distributed iid points there is also a lower bound of the order $\sqrt{d/n}$ for $n \gtrsim d$ which holds with exponentially large (in d) probability, see [28].

Improvements on (5.5) and (5.6), and the construction of points satisfying the latter, seem to be very challenging open problems.

Open Problem 5.3. *Is the upper bound in* (5.6) *sharp for all n that are at most polynomially large in d? In other words, do iid uniform random points have optimal discrepancy in high dimension? Moreover, find explicit deterministic constructions that satisfy $D(\mathcal{P}_n) \lesssim \frac{d^{42}}{\sqrt{n}}$.*

5.5 A machine learning perspective

In this final section we want to provide a different point of view on the setting of this survey. In other literature, especially from data science, it is usually assumed that some "data" $(x_1,y_1),\dots,(x_N,y_N) \in \mathcal{X}\times\mathcal{Y}$ is produced by iid samples of a random vector (X,Y) with distribution ρ on $\mathcal{X}\times\mathcal{Y}$. Our setting (for standard information) amounts to $(X,Y) = (X, f(X))$ for some function $f\colon \mathcal{X} \to \mathcal{Y}$ which we would like to approximate using the given data $(x_1, f(x_1)),\dots,(x_N, f(x_N))$.

In machine learning, one wants to "explain" the data by finding a function $f\colon \mathcal{X} \to \mathcal{Y}$ (a *model*) that maps an input x to an output y. Here, an additional (additive) noise is a typical assumption. Denoting by $\rho_{Y|X}$ the conditional probability distribution given X, the regression function

$$f_\rho(x) = \int_{\mathcal{Y}} y \,\mathrm{d}\rho_{Y|X}(y|x), \quad x \in \mathcal{X},$$

is the best guess of y given x with respect to an L_2-error and the goal is to approximate this function using the given data. For example, one uses empirical best approximation in an hypothesis space $\mathcal{H}$, i.e.,

$$\hat{f}_{\mathbf{z}} := \operatorname*{argmin}_{\hat{f}\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} |\hat{f}(x_i) - y_i|^2,$$

where $\mathbf{z} := ((x_1, y_1), \dots, (x_N, y_N))$ is the given data.

If we define the (squared) *error* of the model g by

$$\mathcal{E}(g) := \mathcal{E}_\rho(g) := \int_{\mathcal{X}} |g(x) - y|^2 \,\mathrm{d}\rho(x, y),$$

then it is easy to verify that the error of the least squares estimator decomposes into

$$\mathcal{E}(\hat{f}_{\mathbf{z}}) = \int_{\mathcal{X}} |\hat{f}_{\mathbf{z}}(x) - f_\rho(x)|^2 \,\mathrm{d}\rho_X(x) + \int_{\mathcal{X}} |f_\rho(x) - y|^2 \,\mathrm{d}\rho(x, y),$$

where ρ_X denotes the marginal of ρ on X.

Naively, our setting in Section 3 corresponds to choosing $\mathcal{H} = V_n$ for a suitable n-dimensional subspace V_n of L_2, where n depends on N, and $f_\rho = f$ (i.e., $y_i = f(x_i)$). This excludes for example the noisy case $(X, Y) = (X, f(X) + \varepsilon)$ where ε is centered noise independent of X.

However, there is another interpretation we would like to comment on. We only know the data $(x_i, y_i)_{i=1}^N \sim \rho$, and the goal is to compute f_ρ. Assuming all x_i are different and there is no noise, we may assume that $y_i = y(x_i)$ for some function $y\colon \mathcal{X} \to \mathcal{Y}$. Hence, we can write $\hat{f}_{\mathbf{z}} = A_N^u(y)$ with A_N^u from (4.2). If the data and hypothesis space V_n is such that $A_N^u(g) = g$ for all $g \in V_n$, and $\hat{f} := \operatorname{argmin}_{g\in V_n} \|f_\rho - g\|_{L_2}$, then

$$\|f_\rho - A_N(y)\|_{L_2} \le \|f_\rho - \hat{f}\|_{L_2} + \|\hat{f} - A_N(y)\|_{L_2} = \mathcal{E}(\hat{f})^{1/2} + \|A_N(y - \hat{f})\|_{L_2}.$$

The first term is often called the *approximation error*, and can not be avoided due to the choice of the hypothesis space. The second is the *sample error*, and depends on the "quality" of the given data. Hence, if we assume that the "error" $\varepsilon = y - \hat{f}$ is not "just random" but has a certain structure, then the considerations of this survey might be of interest for further studies, see e.g. Section 4 of [63].

For the mathematical foundations of learning we refer to [22] and [34].

Acknowledgements

We are grateful for the hospitality of the Leibniz Center for Informatics at Schloss Dagstuhl where they worked on this survey during the Dagstuhl Seminar 23351 *"Algorithms and Complexity for Continuous Problems"*. Further, we thank Albert Cohen, Matthieu Dolbeault, David Krieg and Erich Novak for valuable comments on an earlier version of this survey, as well as for fruitful discussions on the subject in general. We also thank Peter Mathé for pointing out the relation to the Bernstein numbers in Section 3.4. MS has been supported by the Austrian Science Fund (FWF) Project P32405 *"Asymptotic geometric analysis and applications"*.

References

1. Ahlswede R., Winter A.: Strong converse for identification via quantum channels. IEEE Trans. Inform. Theory. **48** (3) 569-579 (2002)
2. Aronszajn N.: Theory of reproducing kernels. Trans. Amer. Math. Soc. **68**, 337-404 (1950)
3. Artstein-Avidan S., Giannopoulos A., Milman V.D.: Asymptotic geometric analysis. Part I, vol. 202 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI (2015)
4. Bach F.: On the Equivalence between Kernel Quadrature Rules and Random Feature Expansions. Journal of Machine Learning Research. **18**, 1-38 (2017)
5. Bandeira A.S., van Handel R.: Sharp nonasymptotic bounds on the norm of random matrices with independent entries. Ann. Probab. **44** (4), 2479-2506 (2016)
6. Bartel F. , Schäfer M., Ullrich T.: Constructive subsampling of finite frames with applications in optimal function recovery. Applied. Comput. Harmon. Anal., to appear.
7. Beck J., Chen W.W.L.: Irregularities of Distribution. Cambridge Tracts in Mathematics. Cambridge University Press. **89** (1987)
8. Bellmann R. Dynamic programming. Princeton University Press (1957)
9. Binev P., Cohen A., Dahmen W., DeVore R., Petrova G., Wojtaszczyk P.: Data Assimilation in Reduced Modeling. SIAM/ASA Journal on Uncertainty Quantification. **5** (1) (2017)
10. Bogachev V.I.: Gaussian measures, vol. 62 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI (1998)
11. Bohn B.: Error analysis of regularized and unregularized least-squares regression on discretized function spaces. Dissertation. Institut für Numerische Simulation, Universität Bonn (2017)
12. Boneh A., Hofri M.: The coupon-collector problem revisited – a survey of engineering problems and computational methods. Stoch. Models. **13** (1), 39-66 (1997)
13. Boyanov B.: Optimal quadrature formulas. Uspekhi Mat. Nauk. **60** (6(366)), 33-52 (2005)
14. Bilyk D., Lacey M., Vagharshakyan A.: On the Small Ball Inequality in all dimensions. J. Funct. Anal. **254** (9), 2470-2502 (2008)
15. Chkifa A., Cohen A., Migliorati G., Nobile F., Tempone R.: Discrete least squares polynomial approximation with random evaluations – application to parametric and stochastic elliptic PDEs. ESAIM: M2AN. **49** (3), 815-837 (2015)
16. Chkifa A., Dolbeault M.: Randomized least-squares with minimal oversampling and interpolation in general spaces. arXiv e-prints. (2023). arXiv: 2306.07435 [math.NA]
17. Cohen A., Dahmen W., DeVore R.A.: Compressed sensing and best k-term approximation. J. Amer. Math. Soc. **22** (1), 211-231 (2009)
18. Cohen A., Davenport M.A., Leviatan D.: On the Stability and Accuracy of Least Squares Approximations. Found. Comput. Math. **13**, 819-834 (2013)
19. Cohen A., DeVore R., Petrova G., Wojtaszczyk P.: Optimal Stable Nonlinear Approximation. Found Comput Math. **22**, 607-648 (2022). doi: 10.1007/s10208-021-09494-z
20. Cohen A., Dolbeault M.: Optimal pointwise sampling for L_2 approximation. J. Complexity. **68**, 101602 (2022)
21. Cohen A., Migliorati G.: Optimal weighted least squares methods. SMAI J. Comput. Math. **3**, 181-203 (2017)
22. Cucker F., Smale S.: On the Mathematical Foundations of Learning. Bulletin of the American Mathematical Society. **39**, 1-49 (2002). doi: 10.1090/S0273-0979-01-00923-5
23. Dai F., Temlyakov V.N.: Sampling Discretization of Integral Norms and Its Application. Proc. Steklov Inst. Math. **319**, 97-109 (2022). doi: 10.1134/S0081543822050091

24. Dai F., Temlyakov V.: Random points are good for universal discretization. J. Math. Anal. Appl. **529** (1), Paper No 127570 (28) (2024)
25. Davidson K.R., Szarek S.J.: Local operator theory, random matrices and Banach spaces. In Handbook of the geometry of Banach spaces, vol. I, pp. 317-366. North-Holland, Amsterdam (2001)
26. DeVore R.A.: Nonlinear approximation. In Acta numerica, 1998, vol. 7 of Acta Numer., pp. 51-150. Cambridge Univ. Press, Cambridge (1998)
27. Dick J., Pillichshammer F.: Discrepancy theory and quasi-Monte Carlo integration. A panorama in Discrepancy Theory, Lecture Notes in Math. Springer. **2107**, (2014)
28. Doerr B.: A lower bound for the discrepancy of a random point set. J. Complexity. **30** (1), 16-20 (2014)
29. Dolbeault M., Krieg D., Ullrich M.: A sharp upper bound for sampling numbers in L_2. Appl. Comput. Harmon. Anal. **63** (2023), doi: 10.1016/j.acha.2022.12.001
30. Donoho D.L.: Compressed sensing. IEEE Trans. Inf. Theory. **52**(4), 1289-1306 (2006). doi: 10.1109/TIT.2006.871582.
31. Dũng D., Temlyakov V.N., Ullrich T.: Hyperbolic Cross Approximation. Advanced Courses in Mathematics – CRM Barcelona. Springer International Publishing (2018)
32. Edmunds D.E., Lang J.: Gelfand numbers and widths. J. Approximation Theory. **166**, 78-84 (2013). doi: 10.1016/j.jat.2012.10.008
33. Erdős P., Renyi A.: On a classical problem of probability theory. Magyar Tud. Akad. Mat. Kutató Int. Közl. **6** (1), 215-220 (1961)
34. Foucart S.: Mathematical pictures at a data science exhibition. Cambridge University Press, Cambridge (2022)
35. Foucart S., Rauhut H.: A mathematical introduction to compressive sensing. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York (2013)
36. Freeman D., Speegle D.: The discretization problem for continuous frames. Advances in Mathematics. **345**, 784-813 (2019)
37. Gal S., Micchelli C.A.: Optimal sequential and non-sequential procedures for evaluating a functional. Appl. Anal. **10**, 105–120 (1980)
38. Garnaev A.Yu., Gluskin E.D.: The widths of a Euclidean ball. Dokl. Akad. Nauk SSSR. **277** (5), 1048-1052 (1984). English translation: Soviet Math. Dokl. **30** (1), 200-204 (1984)
39. Geng J., Wang H.: On the power of standard information for tractability for L_∞ approximation of periodic functions in the worst case setting. J. Complexity. **80**, Paper No 101790 (2024)
40. Giannopoulos A., Milman V.D.: Mean width and diameter of proportional sections of a symmetric convex body, J. Reine Angew. Math. **497**, 113-139 (1998)
41. Giannopoulos A., Milman V.D.: On the diameter of proportional sections of a symmetric convex body. Internat. Math. Res. Notices. **1**, 5-19 (1997)
42. Giannopoulos A., Milman V.D., Tsolomitis A.: Asymptotic formulas for the diameter of sections of symmetric convex bodies. Funct. Anal. **223** (1), 86-108 (2005)
43. Graf S., Luschgy H.: Foundations of quantization for probability distributions. Springer-Verlag, Berlin (2000)
44. Gröchenig K.: Sampling, Marcinkiewicz–Zygmund inequalities, approximation, and quadrature rules. J. Approximation theory. **257**, 105455 (2020)
45. Haberstich C., Nouy A., Perrin G.: Boosted optimal weighted least-squares. Math. Comp. **91** (335), 1281-1315 (2022)
46. Heinrich S.: On the relation between linear n-widths and approximation numbers. J. Approx. Theory. **58** (3), 315-333 (1989)

47. Heinrich S.: Lower bounds for the complexity of Monte Carlo function approximation. J. Complexity. **8** (3), 277-300 (1992)
48. Heinrich S.: Random Approximation in Numerical Analysis. Proceedings of the Conference "Functional Analysis". Essen, 123-171 (1994)
49. Heinrich S.: Randomized complexity of parametric integration and the role of adaption i. finite dimensional case. arXiv e-prints. (2023). arXiv: 2306.13471 [math.NA]
50. Heinrich S., Novak E., Wasilkowski G.W., Woźniakowski H.: The inverse of the star-discrepancy depends linearly on the dimension. Acta Arithmetica. **96**, 279-302 (2001)
51. Hinrichs A.: Covering numbers, Vapnik-Červonenkis classes and bounds for the star-discrepancy. J. Complexity. **20**, 477-483 (2004)
52. Hinrichs A.: Discrepancy, Integration and Tractability. In: Dick J., Kuo F.Y., Peters G.W., Sloan I.H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods. Springer Proceedings in Mathematics & Statistics. **65**, 129-172 (2013)
53. Hinrichs A., Krieg D., Novak E., Prochno J., Ullrich M.: On the power of random information. In: Hickernell F.J., Kritzer P. (eds.) Multivariate Algorithms and Information-Based Complexity, pp. 43-64. DeGruyter, Berlin/Boston (2020)
54. Hinrichs A., Krieg D., Novak E., Prochno J., Ullrich M.: Random sections of ellipsoids and the power of random information. Trans. Am. Math. Soc. **374** (12), 8691-8713 (2021). doi: 10.1090/tran/8502
55. Hinrichs A., Novak E., Ullrich M., Woźniakowski H.: The curse of dimensionality for numerical integration of smooth functions. Math. Comp. **83**, 2853-2863 (2014)
56. Hinrichs A., Novak E., Ullrich M., Woźniakowski H.: The curse of dimensionality for numerical integration of smooth functions II. J. Complexity. **30**, 117-143 (2014)
57. Hinrichs A., Novak E., Ullrich M., Woźniakowski H.: Product rules are optimal for numerical integration in classical smoothness spaces. J. Complexity. **38**, 39-49 (2017)
58. Hinrichs A., Novak E., Vybiral J.: Linear information versus function evaluations for L_2-approximation. J. Approx. Theory. **153** (1), 97-107 (2008)
59. Hinrichs A., Prochno J., Sonnleitner M.: Random sections of ℓ_p-ellipsoids, optimal recovery and Gelfand numbers of diagonal operators. J. Approx. Theory. **293**, Paper No 105919 (2023)
60. Hlawka E.: Funktionen von beschränkter Variation in der Theorie der Gleichverteilung. Ann. Math. Pura Appl. **54**, 325-333 (1961). doi: 10.1007/s10208-020-09463-y
61. Jahn T., Ullrich T., Voigtlaender F.: Sampling numbers of smoothness classes via ℓ^1-minimization. J. Complexity. **79**, Paper No 101786 (35) (2023)
62. Kashin B.S.: The widths of certain finite-dimensional sets and classes of smooth functions. Izv. Akad. Nauk SSSR Ser. Mat. **41** (2), 334-351 (1977). English translation: Math. USSR-Izv. **11** (2), 317-333 (1978)
63. Kashin B., Kosov E., Limonova I., Temlyakov V.: Sampling discretization and related problems. J. Complexity. 101653 (2022). doi: 10.1016/j.jco.2022.101653
64. Krieg D.: Optimal Monte Carlo Methods for L_2-Approximation. Constr. Approx. **49**, 385-403 (2019). doi: 10.1007/s00365-018-9428-4
65. Krieg D.: Uniform recovery of high-dimensional C^r-functions. J.Ċomplexity. **50**, 116-126 (2019)
66. Krieg D.: Tractability of sampling recovery on unweighted function classes. arXiv e-prints. (2023). arXiv: 2304.14169 [math.NA]
67. Krieg D., Novak E., Sonnleitner M.: Recovery of Sobolev functions restricted to iid sampling. Math. Comp. **91**, 2715-2738 (2022)
68. Krieg D., Pozharska K., Ullrich M., Ullrich T.: Sampling recovery in L_2 and other norms. arXiv e-prints. (2023). arXiv: 2305.07539 [math.NA]

69. Krieg D., Siedlecki P., Ullrich M., Woźniakowski H.: Exponential tractability of L_2-approximation with function values. Adv. Comput. Math. **49**, Article number 18 (2023)
70. Krieg D., Sonnleitner M.: Random points are optimal for the approximation of Sobolev functions. IMA J. Numer. Anal. (2023)
71. Krieg D., Sonnleitner M.: Function recovery on manifolds using scattered data. arXiv e-prints. (2021). arXiv: 2109.04106 [math.NA]
72. Krieg D., Ullrich M.: Function values are enough for L_2-approximation. Found. Comput. Math. **21**, 1141-1151 (2021). doi: 10.1007/s10208-020-09481-w
73. Krieg D., Ullrich M.: Function values are enough for L_2-approximation: Part II. J. Complexity. **66**, (2021). doi: 10.1016/j.jco.2021.101569
74. Krieg D., Vybiral J.: New Lower Bounds for the Integration of Periodic Functions. J. Fourier Anal Appl. **29** (41), (2023). doi: 10.1007/s00041-023-10021-7
75. Kunsch R., Novak E., Wnuk M.: Randomized approximation of summable sequences – adaptive and non-adaptive. arXiv e-prints. (2023). arXiv:2308.01705 [math.NA]
76. Kuo F., Wasilkowski G.W., Woźniakowski H.: On the power of standard information for multivariate approximation in the worst case setting. J. Approx. Theory. **158**, 97-125 (2009)
77. Laurent B., Massart P.: Adaptive estimation of a quadratic functional by model selection. Ann. Statist. **28** (5), 1302-1338 (2000)
78. Limonova I., Temlyakov V.: On sampling discretization in L_2. J. Math. Anal. Appl. **515** (2), 126457 (2022)
79. Litvak A.E.: A remark on the minimal dispersion. Commun. Contemp. Math. **23** (6), Paper No 2050060 (18) (2021)
80. Litvak A.E., Pajor A., Tomczak-Jaegermann N.: Diameters of sections and coverings of convex bodies. J . Funct. Anal. **231** (2), 438-457 (2006)
81. Litvak A.E., Tomczak-Jaegermann N.: Random aspects of high-dimensional convex bodies. GAFA, Lecture Notes in Math. Springer-Verlag. **1745**, 169-190 (2000)
82. Marcus A., Spielman D., Srivastava N.: Interlacing families II: Mixed characteristic polynomials and the Kadison-Singer problem. Ann. of Math. **182**, 327-350 (2015). doi: 10.4007 /annals.2015.182.1.8.
83. Mathé P.: Random approximation of Sobolev embeddings. J. Complexity. **7** (3), 261-281 (1991)
84. Maz'ja V.G.: Sobolev spaces. Springer-Verlag, Berlin (1985)
85. Meaney C.: Remarks on the Rademacher-Menshov theorem. Proc. Centre Math. Appl. Austral. Nat. Univ. **42**, 100-110 (2007)
86. Mendelson S., Pajor A.: On singular values of matrices with independent rows. Bernoulli. **12** (5), 761-773 (2006)
87. Nagel N., Schäfer M., Ullrich T.: A new upper bound for sampling numbers. Found. Comput. Math. **22** (2), 445-468 (2022)
88. Narcowich F.J., Ward J.D., Wendland H.: Sobolev bounds on functions with scattered zeros, with applicationsc to radial basis function surface fitting. Math. Comp. **74** (250), 743-763 (2004)
89. Nitzan S., Olevskii A., Ulanovskii A.: Exponential frames on unbounded sets. Proc. Amer. Math. Soc. **144** (1), 109-118 (2016). doi: 10.1090/proc/12868
90. Novak E.: Deterministic and Stochastic Error Bounds in Numerical Analysis. Lecture Notes in Math. Springer. **1349**, (1988)
91. Novak E.: Some Results on the Complexity of Numerical Integration. In: Monte Carlo and Quasi-Monte Carlo Methods. Springer Proceedings in Math. & Statistics. **163**, 161-183 (2016)
92. Novak E., Triebel H.: Function spaces in Lipschitz domains and optimal rates of convergence for sampling. Constr. Approx. **23**, 325-350 (2006)

93. Novak E., Woźniakowski H.: Approximation of infinitely differentiable multivariate functions is intractable. J. Complexity. **25**, 398-404 (2009)
94. Novak E., Woźniakowski H.: On the power of function values for the approximation problem in various settings. Surveys in Approximation Theory. **6**, 1-23 (2011)
95. Novak E., Woźniakowski H.: Tractability of multivariate problems. Volume I: Linear Information. European Mathematical Society, Zürich (2008)
96. Novak E., Woźniakowski H.: Tractability of multivariate problems. Volume II: Standard Information for Functionals. European Mathematical Society (EMS), Zürich (2010)
97. Novak E., Woźniakowski H.: Tractability of multivariate problems. Volume III: Standard Information for Operators. European Mathematical Society (EMS), Zürich (2012)
98. Oliveira R.I.: Sums of random Hermitian matrices and an inequality by Rudelson. Electr. Comm. Probab. **15**, 203-212 (2010)
99. Pagès G.: A space quantization method for numerical integration. J. Comput. Appl. Math. **89** (1), 1-38 (1998)
100. Pietsch A.: s-numbers of operators in Banach spaces. Studia Math. **51**, 201-223 (1974)
101. Pietsch A.: Eigenvalues and s-numbers, vol. 13 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge (1987)
102. Pinkus A.: n-widths in approximation theory. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics (MATHE3, vol. 7), Springer Berlin, Heidelberg (1985). doi: 10.1007/978-3-642-69894-1
103. Reznikov A., Saff E.B.: The covering radius of randomly distributed points on a manifold. Int. Math. Res. Not. IMRN. **2016** (19), 6065-6094 (2016)
104. Rudelson M., Vershynin R.: Smallest singular value of a random rectangular matrix. Comm. Pure Appl. Math. **62** (12), 1707-1739 (2009)
105. Salem R.: A new proof of a theorem of Menchoff. Duke Mathematical Journal. **8** (2), 269-272 (1941)
106. Sloan I.H., Woźniakowski H.: When Are Quasi-Monte Carlo Algorithms Efficient for High Dimensional Integrals? J. Complexity. **14**, 1-33 (1998)
107. Sonnleitner M.: The power of random information for numerical approximation and integration. PhD thesis, University of Passau (2022). arXiv: 2209.07266
108. Steinwart I., Scovel C.: Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs. Constr. Approx. **35** (3), 363-417 (2012)
109. Sukharev A.G.: Optimal numerical integration formulas for some classes of functions of several variables. Soviet Math. Dokl. **20**, 472-475 (1979)
110. Talagrand M.: Sharper bounds for Gaussian and empirical processes. Ann. Probab. **22** (1), 28-76 (1994)
111. Temlyakov V.N.: Multivariate Approximation, vol. 32 of Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press (2018)
112. Temlyakov V.N.: On optimal recovery in L_2. J. Complexity. **65**, 101545 (2020). doi: 10.1016/j.jco.2020.101545
113. Traub J., Wasilkowski G., Woźniakowski H.: Information-Based Complexity. Acad. Press (1988)
114. Tropp J.: User-Friendly tail bounds for sums of random matrices. Foundations of Computational mathematics. textbf12, 389-434 (2012)
115. Ullrich M.: On the worst-case error of least squares algorithms for L_2-approximation with high probability. J. Complexity. **60** (2020). doi: 10.1016/j.jco.2020.101484
116. Ullrich M., Vybíral J.: An upper bound on the minimal dispersion. J. Complexity. **45**, 120-126 (2018). doi: 10.1016/j.jco.2017.11.003

117. Wasilkowski G.W., Woźniakowski H.: On the power of standard information for weighted approximation. Found. Comput. Math. **1** (4), 417-434 (2001)
118. Wasilkowski G.W., Woźniakowski H.: The power of standard information for multivariate approximation in the randomized setting. Math. Comp. **76** (258), 965-988 (2007)
119. Wendland H.: Scattered data approximation. Cambridge University Press, Cambridge (2005)

Received 06.10.2023
Revised 13.11.2023