L_2-approximation based on Gaussian information, function values or other information

Mario Ullrich
JKU Linz & MCFAM Moscow

Sampling recovery workshop
Online, May 2021

Motivation

We want to recover/approximate

$$\text{a function } f\colon D \to \mathbb{R}$$

(or some property of it) up to

$$\text{a certain error } \varepsilon > 0,$$

where f is only known through

some pieces of information.

Motivation

We want to recover/approximate

a function $f\colon D \to \mathbb{R}$?

(or some property of it) up to

a certain error $\varepsilon > 0$, ??

where f is only known through

some pieces of information. ???

During this talk ...

we consider

- a measure space $(D, \mathcal{A}, \mu)$,
- $L_2 = L_2(D, \mathcal{A}, \mu)$: the square-integrable functions w.r.t. μ, and
- a separable metric space $F \hookrightarrow L_2$ of functions on D.

For example:

- $D = [0,1]^d$ or $D = \mathbb{R}^d$ or $D = \mathbb{N}$, with arbitrary μ, and
- F is the unit ball of a separable normed space.

($F \hookrightarrow L_2$ means here that $\mathrm{id}\colon F \to L_2$, $\mathrm{id}(f) = f$, is injective and compact.)

Approximation

We want to "compute" an L_2-approximation of $f \in F$ based on a finite (preferably small) number of information, because we ...

- don't know f and we can only take some measurements, or
- know f, but want to compress it because of computing issues.

What information is allowed,

and how important is this choice?

(The statement "$f \in F$" can be seen as the a priori knowledge about f.)

Information

Information of a function $f \in F$ is given by $L(f)$ for some linear functional $L \colon F \to \mathbb{R}$.

In general, we do not have access to arbitrary $L \in F'$ (=dual of F).

Instead, we have a class of **admissible information** $\Lambda \subset F'$, e.g.,

- certain expectations of f,
- coefficients w.r.t. a given basis,
- **function values:** $f(x)$ for $x \in D$.

Algorithms & error

For information (maps) $L_1, \dots, L_n \in \Lambda$, we study **linear algorithms**:

$$A_n(f) \;=\; \sum_{i=1}^{n} L_i(f) \cdot \varphi_i$$

for some $\varphi_i \in L_2$. So, A_n is specified by L_i, φ_i.

We want to bound the **worst-case error** over F:

$$e(A_n, F) \;=\; \sup_{f \in F} \left\| f - A_n(f) \right\|_{L_2}.$$

(Several other settings are possible here. Linearity has advantages.)

Minimal worst-case errors

We are interested in the **(linear) sampling numbers**

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

i.e., the minimal error that can be achieved with n function values.

As a benchmark, we use the **approximation numbers** (linear width)

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

i.e., the minimal error that can be achieved with arbitrary info.

How good are function values?

The a_n's are well understood, but the g_n's are harder to analyze.

We clearly have

$$a_n(F) \leq g_n(F)$$

if point evaluation $f \mapsto f(x)$ is a continuous linear functional on F.

How large is the difference between g_n and a_n?

Earlier results

Several specific, but only some general bounds were known before.

A negative result [Hinrichs/Novak/Vybíral 2008]

For any $(a_n) \notin \ell_2$, there exist F with $a_n(F) = a_n$ for all n, but

$$g_n(F) \geq \frac{1}{\log\log(n)}.$$

for infinitely many n.

A positive result [Kuo/Wasilkowski/Woźniakowski 2009]

For unit balls of Hilbert spaces H with $a_n(H) \lesssim n^{-\alpha}$, $\alpha > 1/2$, we have

$$g_n(H) \lesssim n^{-\alpha \frac{2\alpha}{2\alpha+1}} \lesssim n^{-\alpha/2}.$$

A very positive result

We now have this general result on the **power of function values**.

Theorem [Krieg/U 2019; U 2020; Krieg/U 2021]

Let $F \hookrightarrow L_2$ be a separable metric space of functions on D, such that point evaluation is continuous on F.

Then, for every $0 < p < 2$, there is a constant $c_p > 0$, depending only on p, such that, for all $n \geq 2$, we have

$$g_N(F) \leq \sqrt{\log n} \left(\frac{1}{n} \sum_{k \geq n} a_k(F)^p \right)^{1/p}$$

for $N \geq c_p \cdot n$.

For unit balls of Hilbert spaces, $p = 2$ also works.[Nagel, Schäfer, T. Ullrich, 2020]

In particular, ...

Corollary

If F is such that

$$a_n(F) \lesssim n^{-\alpha} \log^{\beta}(n)$$

for some $\alpha > 1/2$ and $\beta \in \mathbb{R}$, then we obtain

$$g_n(F) \lesssim n^{-\alpha} \log^{\beta+1/2}(n).$$

Stated differently: If $n \approx (\frac{1}{\varepsilon})^q$, $q < 2$, (arbitrary) infos are enough for an approximation with error $\varepsilon > 0$, then $\left(\frac{\sqrt{\log(1/\varepsilon)}}{\varepsilon}\right)^q$ function values can do the same.

Original motivation

However, our original motivation was different. We wanted to know:

How special is optimal information?

To be precise, let us start with a discussion of optimal information.

In what follows, we use the notation

- F – separable metric space
- H – unit ball of a Hilbert space

Hilbert spaces: Singular value decomposition

The $a_n(H)$'s can be given (in theory) using the SVD:

If $\mathrm{id}\colon H \to L_2$ is compact, there is an

$$\textbf{orthogonal basis} \quad \mathcal{B} = \{b_k \colon k \in \mathbb{N}\} \text{ of } H$$

that consists of eigenfunctions of $\mathrm{id}^* \cdot \mathrm{id}\colon H \to H$. We have that

- $\mathcal{B}$ is also orthogonal in L_2, and
- we assume $\|b_j\|_{L_2} = 1$, and $\|b_1\|_H \leq \|b_2\|_H \leq \dots$

Then,

$$a_n(H) = \frac{1}{\|b_{n+1}\|_H}.$$

Optimal algorithm: projection

Using this notation, we have that

$$f = \sum_{j=1}^{\infty} \langle f, b_j \rangle_{L_2} \, b_j = \sum_{j=1}^{\infty} \frac{\langle f, b_j \rangle_H}{\langle b_j, b_j \rangle_H} \cdot b_j$$

converges in H for every $f \in H$.

The optimal algorithm based on n linear functionals is given by

$$P_n(f) := \sum_{j \le n} \langle f, b_j \rangle_{L_2} \, b_j,$$

which is the orthogonal projection onto

$$V_n := \mathrm{span}\{b_1, \dots, b_n\}.$$

Optimal algorithm: error

We obtain that

$$P_n(f) = \sum_{j\leq n} \langle f, b_j \rangle_{L_2} \, b_j$$

satisfies

$$a_n(H) = \sup_{f\in H\colon \|f\|_H \leq 1} \big\| f - P_n(f) \big\|_{L_2} = \frac{\|b_{n+1}\|_{L_2}}{\|b_{n+1}\|_H} = \frac{1}{\|b_{n+1}\|_H}.$$

General classes: A "good" basis

It is not hard to show that similar holds true for general classes F:

Lemma

There is an orthonormal system $\{b_k : k \in \mathbb{N}\}$ in L_2 such that the orthogonal projection P_n onto the span $V_n = \mathrm{span}\{b_1, \dots, b_n\}$ satisfies

$$\sup_{f \in F} \|f - P_n f\|_{L_2} \leq 2\, a_{n/4}(F), \qquad n \in \mathbb{N}.$$

- This system is not known in general.
- The '$n/4$' might be problematic for rapidly decaying a_n.
- From now on, $\{b_k\}$ will always be as above.

Random information

Our attempt to study the "rarity" of optimal info was to ask:

How good is random information?

Recall that we are in the worst-case setting:

For given info, there is no randomness.

Fixed information

To study "random" information, we first introduce

$$e(F, N_n) \;:=\; \inf_{\varphi_1,\dots,\varphi_n\in L_2}\; \sup_{f\in F} \left\| f - \sum_{i=1}^{n} L_i(f)\,\varphi_i \right\|_{L_2},$$

i.e., the minimal error that can be achieved by linear algorithms based on the **fixed info**

$$N_n(f) := \Big(L_1(f), \dots, L_n(f)\Big).$$

Clearly,

$$a_n(F) \;=\; \inf_{N_n\in(F')^n} e(F, N_n)$$

What is a good model for random info?

In the 'simple' examples $F \subset \mathbb{R}^m$, $m \in \mathbb{N}$, it might be natural to consider uniformly distributed info from the sphere

$$L_i(f) = \langle f, y^{(i)} \rangle_2, \quad \text{where} \quad y^{(i)} \overset{\text{iid}}{\sim} \mathbb{S}^{m-1}.$$

Equivalently, we can consider **Gaussian information**

$$L_i(f) = \sum_{j=1}^{m} g_{ij} f_j, \quad \text{where} \quad g_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

The latter makes also sense for $m = \infty$.

A geometric formulation ($m < \infty$)

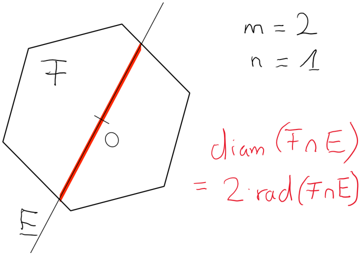
Assume that $F \subset \mathbb{R}^m$ is convex and symmetric. Then

$$e(F, N_n) \;=\; \sup\Big\{ \|f\|_2 : \; f \in F, \; N_n(f) = 0 \Big\}.$$

In other words,

$$e(F, N_n) \;=\; \mathrm{rad}(F \cap E),$$

i.e., the radius of the intersection with a hyperplane $E \subset \mathbb{R}^m$ with codimension n (uniformly distributed on the Grassmannian).



Ellipsoids aka. Hilbert spaces

For $1 = \sigma_1 \geq \sigma_2 \geq \ldots \geq 0$ and $n < m$, consider

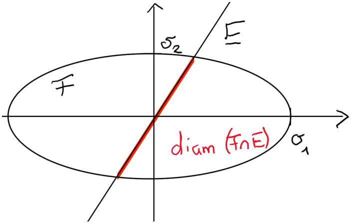
$$H = \left\{ f = (f_1, \ldots, f_m) \in \mathbb{R}^m : \sum_{j=1}^{m} \left(\frac{f_j}{\sigma_j} \right)^2 \leq 1 \right\}.$$

Optimal information is given by $N_n^*(f) = (f_1, \ldots, f_n)$ and

$$a_n(H) = e(H, N_n^*) = \sigma_{n+1}.$$

How good is Gaussian information

$$N_n(f) = (L_1(f), \ldots, L_n(f)) \quad ?$$



To ease the presentation, we stick to the case $m = \infty$.

Gaussian info might be useless!

Theorem [Hinrichs/Krieg/Novak/Prochno/U 2018]

If $\sigma \notin \ell_2$, then, for Gaussian info N_n, we almost surely have

$$e(H, N_n) \,=\, \sigma_1.$$

Proof: Let $\varepsilon > 0$.

- A result of Kahane (1985) implies that $N_n(H) = \mathbb{R}^n$ a.s.
- In particular, there is $y \in H$ with $N_n y = \frac{\sigma_1(1-\varepsilon)}{\varepsilon} N_n e_1$.
- Then $x = \sigma_1(1-\varepsilon)e_1 - \varepsilon y \in F$ with $N_n x = 0$ and

$$\|x\|_2 \,\geq\, x_1 \,\geq\, \sigma_1(1-2\varepsilon).$$

- Since $\pm x$ cannot be distinguished, $e(H, N_n) \geq \sigma_1(1-2\varepsilon)$.

Gaussian info might be optimal!

Theorem [Hinrichs/Krieg/Novak/Prochno/U 2018]

Let $\sigma \in \ell_2$. Then, for Gaussian info N_n, we have that

$$e(H, N_n) \leq \sqrt{\frac{C}{n} \sum_{j>cn} \sigma_j^2}.$$

with probability at least $1 - e^{-cn}$ for some absolute constants c, C.

This is achieved by the **algorithm** $A_n = G^+ \circ N_n$, where G^+ is the Moore-Penrose-inverse of $G = (g_{ij})_{i \leq n, j \leq k}$ and $k = n/2$.

Note that $G = N_n|_{\mathbb{R}^k}$.

Proof of the upper bound

Since $A_n = G^+ N_n$ with $G = N_n|_{\mathbb{R}^k}$, we have that $A_n(f) = f$ for $f \in \mathbb{R}^k$, if G has full rank. This holds with probability 1.

Then, for $f \in F$, let $P_k(f)$ be the projection to $\mathbb{R}^k$. We have

$$\|f - A_n(f)\|_2 \;\leq\; \|f - P_k(f)\|_2 + \|A_n(f) - P_k(f)\|_2.$$

The first term is bounded by σ_{k+1}. The second term satisfies

$$A_n(f) - P_k(f) = A_n(f - P_k(f)) = G^+ \Gamma z,$$

with $z = \left(\frac{f_j}{\sigma_j}\right)_{j>k}$ and $\Gamma = (\sigma_j g_{ij})_{i \leq n, j>k} \in \mathbb{R}^{n \times \infty}$. Since $\|z\|_2 \leq 1$,

$$\|A_n(f) - P_k(f)\|_2 \;\leq\; \|G^+ \colon \ell_2^n \to \ell_2^k\| \cdot \|\Gamma \colon \ell_2 \to \ell_2^n\|.$$

Proof of the upper bound II

We have, for $f \in F$, that

$$\|f - A_n(f)\|_2 \;\leq\; \sigma_{k+1} \;+\; \|G^+\colon \ell_2^n \to \ell_2^k\| \cdot \|\Gamma\colon \ell_2 \to \ell_2^n\|.$$

The norm of G^+ is the inverse of the smallest singular value of G and roughly $n^{-1/2}$. The norm of $\Gamma = (\sigma_j g_{ij})_{i \leq n, j > k}$ is roughly

$$n^{1/2} \max\Big\{ \Big(\frac{1}{k}\sum_{j>k} \sigma_j^2\Big)^{1/2}, \sigma_{k+1} \Big\}.$$

See e.g. [Davidson/Szarek 2001, Bandeira/Van Handel 2016].

(Note that G and Γ are independent random matrices.)

Power of Gaussian information

Recall that $H = \left\{ f = (f_1, f_2, \dots) \in \ell_2 :\ \sum_{j=1}^{\infty} \left(\frac{f_j}{\sigma_j}\right)^2 \le 1 \right\}$.

For sequences (σ_j) of **polynomial decay**, we obtain the following.

Theorem [Hinrichs/Krieg/Novak/Prochno/U 2018]

Let $\sigma_n \asymp n^{-\alpha} \log^{\beta} n$ for some $\alpha > 0$ and $\beta \in \mathbb{R}$.

Then, for Gaussian info N_n, and with $a_n := a_n(H) = \sigma_{n+1}$, we have

$$\mathbb{E}\Big[e(H, N_n)\Big] \asymp \begin{cases} a_0 \ (= \sigma_1) & \text{for} \quad \sigma \notin \ell_2, \\ a_n & \text{for} \quad \alpha > 1/2, \\ a_n \sqrt{\log n} & \text{else.} \end{cases}$$

Analogous estimates hold with high probability.

How special is optimal information?

Although this is a very special setting, one may deduce the following heuristic:

1. For $(a_n) \notin \ell_2$: Optimal information is rare.
2. For $(a_n) \in \ell_2$: (Almost) optimal information is nothing special.

Does the latter imply that one can

restrict to smaller classes of information,

maybe even for more general problem classes?

Function values

Recall the similar scenario for approximation using function values.

A negative result [Hinrichs/Novak/Vybíral 2008]

For any $(a_n) \notin \ell_2$, there exist F with $a_n(F) = a_n$ for all n, but

$$g_n(F) \geq \frac{1}{\log\log(n)}.$$

for infinitely many n.

A positive result [Kuo/Wasilkowski/Woźniakowski 2009]

For unit balls of Hilbert spaces H with $a_n(H) \lesssim n^{-\alpha}$, $\alpha > 1/2$, we have

$$g_n(H) \lesssim n^{-\alpha \frac{2\alpha}{2\alpha+1}} \lesssim n^{-\alpha/2}.$$

Generalization

In order to generalize the methods from above to general F, let

- $\{b_k\colon k \in \mathbb{N}\}$ be a "good" basis for $F \subset \mathbb{R}^D$,
- P_n be the orthogonal projection onto $V_n = \mathrm{span}\{b_1, \ldots, b_n\}$,
- $N(f) = \big(L_1(f), \ldots, L_N(f)\big)$, $N \in \mathbb{N}$ (and $N\colon F \to \mathbb{R}^N$),
- $G = (L_i(b_j))_{i \le N, j \le n} \in \mathbb{R}^{N \times n}$, (i.e., $G \cong N|_{V_n}$)
- the algorithm

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

Least squares

Note that this algorithm is a **least squares estimator**:

If G has full rank, then

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

It is linear and **exact on** V_n.

See the talk of Karlheinz & Albert for introduction and discussion.

Least squares for function values

It is a classical for $L_i(f) = f(x_i)$, $x_i \in D$, to study **weighted least squares methods**:

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

for some weigths $d_i > 0$, $x_i \in D$ and $V_n = \mathrm{span}\{b_1, \ldots, b_n\} \subset L_2$.

The analysis often boils down to the study of quantities depending on

$$\sum_{k=1}^{n} |b_k(x)|^2 \qquad \text{and} \qquad (f - P_n f)(x).$$

There are many approaches: See talks of Albert, Tino and Volodya.

Least squares: our approach

To compare $g_n(F)$ and $a_n(F)$, we consider

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

with $\varrho\colon D \to \mathbb{R}$,

$$\varrho(x) \ := \ \frac{1}{2} \left(\frac{1}{n} \sum_{k \le n} |b_k(x)|^2 + \sum_{k > n} w_k \, |b_k(x)|^2 \right)$$

for some sequence (w_k), s.t. ρ is a μ-density, and choose

$$x_1, \dots, x_N \overset{\text{iid}}{\sim} \rho \cdot \mathrm{d}\mu.$$

The general result

Theorem [Krieg/U 2021]

Let $F_0 \subset L_2(\mu)$ be a countable set and $x_1, \ldots, x_N \overset{\text{iid}}{\sim} \rho \cdot \mathrm{d}\mu$.

Then, for every $0 < p < 2$, there is a constant $c_p > 0$, depending only on p, such that, for all $n \geq 2$, we have

$$e(A_N, F_0) \leq \left(\frac{1}{n} \sum_{k \geq n} a_k(F_0)^p \right)^{1/p}$$

for $N \geq c_p \, n \log(n)$ with probability at least $1 - \frac{1}{n^2}$.

(For unit balls of Hilbert spaces, $p = 2$ also works.[Krieg/U 2019])

The proof

The first important insight is that A_N can be written as

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

where $N\colon F_0 \to \mathbb{R}^n$ with $N(f) = \Big(\varrho(x_i)^{-1/2} f(x_i)\Big)_{i \le N}$ is the **weighted information mapping** and

$G^+ \in \mathbb{R}^{n \times N}$ is the Moore-Penrose inverse of the matrix

$$G = \left(\frac{b_j(x_i)}{\sqrt{\varrho(x_i)}}\right)_{i \le N, j \le n} \in \mathbb{R}^{N \times n}.$$

The proof II

Again, since A_N is exact on V_n, we obtain

$$\begin{aligned}\|f - A_N f\|_{L_2} &\leq \|f - P_n f\|_{L_2} + \|P_n f - A_n f\|_{L_2} \\ &\leq a_n + \left\|G^+ N(f - P_n f)\right\|_{\ell_2^n} \\ &\leq a_n + \left\|G^+ \colon \ell_2^N \to \ell_2^n\right\| \cdot \|N(f - P_n f)\|_{\ell_2^N}\end{aligned}$$

and hence

$$\begin{aligned}e(A_N, F_0) &= \sup_{f \in F_0} \|f - A_N(f)\|_{L_2} \\ &\leq a_n + s_{\min}(G)^{-1} \sup_{f \in F_0} \|N(f - P_n f)\|_{\ell_2^N},\end{aligned}$$

where $s_{\min}$ denotes the smallest singular value.

The proof III

$$e(A_N, F_0) \;\leq\; a_n + s_{\min}(G)^{-1} \sup_{f \in F_0} \|N(f - P_n f)\|_{\ell_2^N},$$

We will show that

Fact 1: $s_{\min}(G\colon \ell_2^n \to \ell_2^N) \;\gtrsim\; \sqrt{N}$

Fact 2: $\sup_{f \in F_0} \|N(f - P_n f)\|_{\ell_2^N} \;\lesssim\; \sqrt{n \log n} \left(\frac{1}{n} \sum_{k \geq n} a_k^p \right)^{1/p}$

for $N \approx c_p\, n \log(n)$ simultaneously with high probability.

The proof: main tool

Proposition [Oliveira 2010, Mendelson/Pajor 2006]

Let X be a random vector in $\mathbb{C}^k$ with $\|X\|_2 \leq R$ with probability 1, and let $X_1, X_2, \ldots$ be independent copies of X. Additionally, let $E := \mathbb{E}(XX^*)$ satisfy $\|E\| \leq 1$, where $\|E\|$ denotes the spectral norm of E. Then, for all $t \geq \frac{1}{2}$,

$$\mathbb{P}\left(\left\| \sum_{i=1}^{N} X_i X_i^* - N \cdot E \right\| \geq N \cdot t \right) \leq 4N^2 \exp\left(-\frac{N}{32R^2} t\right).$$

Note that the bound is dimension-free.

The proof of Fact 1

Let $X_i := \varrho(x_i)^{-1/2}(b_1(x_i), \dots, b_n(x_i))^\top$ with $x_i \sim \rho$. Then, we have

$$\sum_{i=1}^{N} X_i X_i^* \;=\; G^* G \;=\; \left(\sum_{i=1}^{N} \frac{\overline{b_j(x_i)}\, b_k(x_i)}{\varrho(x_i)} \right)_{j,k\le n} \in \mathbb{R}^{n\times n}$$

and $E = \mathbb{E}(XX^*) = \mathrm{diag}(1, \dots, 1)$, i.e., $\|E\| = 1$. Moreover,

$$\|X_i\|_2^2 \;=\; \varrho(x_i)^{-1} \sum_{k\le n} |b_k(x_i)|^2 \;\le\; 2n \;=:\; R^2,$$

since

$$\varrho(x) \;\ge\; \frac{1}{2n} \sum_{k\le n} |b_k(x)|^2.$$

The proof of Fact 1

With $t = \frac{1}{2}$ and $N = \lceil C_1 n \log n \rceil$, we obtain

$$\mathbb{P}\Big(\|G^*G - NE\| \geq \frac{N}{2}\Big) \;\leq\; \frac{4}{n^2}$$

if the constant $C_1 > 0$ is large enough. We obtain

$$s_{\min}(G)^2 \;=\; s_{\min}(G^*G) \;\geq\; s_{\min}(NE) - \|G^*G - NE\| \;\geq\; \frac{N}{2}$$

with probability at least $1 - \frac{4}{n^2}$.

The proof of Fact 2: Decomposition

With $I_\ell := \{n2^\ell + 1, \ldots, n2^{\ell+1}\}$, $\ell \geq 0$, and the random matrices

$$\Gamma_\ell := \Big(\varrho(x_i)^{-1/2} b_k(x_i)\Big)_{i \leq N, k \in I_\ell} \in \mathbb{R}^{N \times n2^\ell},$$

and $\hat{f}_\ell := (\langle f, b_k \rangle_{L_2})_{k \in I_\ell}$, we obtain that

$$\|N(f - P_n f)\|_{\ell_2^N} \overset{?}{=} \Big\| \sum_{\ell=0}^{\infty} \Gamma_\ell \hat{f}_\ell \Big\|_{\ell_2^N} \leq \sum_{\ell=0}^{\infty} \|\Gamma_\ell \colon \ell_2(I_\ell) \to \ell_2^m\| \, \|\hat{f}_\ell\|_{\ell_2(I_\ell)}$$

$$\leq 2 \sum_{\ell=0}^{\infty} \|\Gamma_\ell \colon \ell_2(I_\ell) \to \ell_2^m\| \;\; a_{n2^{\ell-2}}(F_0)$$

for all $f \in F_0$.

The proof of Fact 2: individual blocks

For fixed ℓ, let $X_i := \varrho(x_i)^{-1/2}(b_k(x_i))_{k\in I_\ell}^\top$ with $x_i \sim \rho$. We have

$$\sum_{i=1}^N X_i X_i^* = \Gamma_\ell^* \Gamma_\ell = \left(\sum_{i=1}^N \frac{\overline{b_j(x_i)}\, b_k(x_i)}{\varrho(x_i)}\right)_{j,k\in I_\ell} \in \mathbb{R}^{n2^\ell \times n2^\ell}$$

and $E = \mathbb{E}(XX^*) = \mathrm{diag}(1,\dots,1)$, i.e., $\|E\| = 1$. Moreover,

$$\|X_i\|_2^2 = \varrho(x_i)^{-1} \sum_{k\in I_\ell} |b_k(x_i)|^2 \le \frac{2}{w_{n2^{\ell+1}}} =: R^2,$$

since

$$\varrho(x) \ge \frac{1}{2} \sum_{k\in I_\ell} w_k |b_k(x)|^2 \ge \frac{w_{n2^{\ell+1}}}{2} \sum_{k\in I_\ell} |b_k(x)|^2.$$

The proof of Fact 2: union bound

With $t \approx \frac{\log(n\ell)}{w_{n2^\ell} \log(n)}$ and $N = \lceil C_1 n \log n \rceil$, we obtain with $\|\Gamma_\ell\|^2 \leq m + \|\Gamma_\ell^* \Gamma_\ell - mE\|$ that

$$\mathbb{P}\left(\|\Gamma_\ell\|^2 \geq C_2\, n \log(n)\, B_\ell^2\right) \leq \frac{4}{n^2(\ell+1)^2\pi^2}$$

for some $B_\ell \gg \sqrt{\ell\, 2^\ell}$ that is independent of n, N.

We obtain by a union bound that

$$\mathbb{P}\left(\exists \ell \in \mathbb{N}_0 \colon \|\Gamma_\ell\|^2 \geq C_2\, n \log(n)\, B_\ell^2\right) \leq \frac{1}{n^2}.$$

The proof of Fact 2: some calculation

Hence,

$$\|N(f - P_n f)\|_{\ell_2^N} \lesssim n \log(n) \sum_{\ell=0}^{\infty} B_\ell \, a_{n2^\ell}(F_0)$$

for all $f \in F_0$ with probability at least $1 - \frac{1}{n^2}$.

Monotonicity of (a_n) gives

$$\sum_{k \geq n} a_k^p \geq n(2^\ell - 1) \, a_{n2^\ell}^p$$

for $\ell \geq 1$ and thus $a_{n2^\ell} \lesssim 2^{-\ell/p} \left(\frac{1}{n} \sum_{k \geq n} a_k^p \right)^{1/p}$.

We can choose suitable w_k, B_ℓ if $p \in (0, 2)$, which finishes the proof.

The proof of Fact 2: point-wise convergence

It remains to verify $\|N(f - P_n f)\|_{\ell_2^N} \overset{?}{=} \left\| \sum_{\ell=0}^{\infty} \Gamma_\ell \hat{f}_\ell \right\|_{\ell_2^N}$:

We implicitly use

$$(f - P_n f)(x_i) = \sum_{k>n} \hat{f}(k)\, b_k(x_i).$$

Rademacher-Menchov theorem

Let F_0 be **countable** with $\left(\sqrt{\frac{\log(k)}{k}} \cdot a_k(F_0)\right) \in \ell_2$. Then, there is a measurable subset D_0 of D with $\mu(D \setminus D_0) = 0$ such that

$$f(x) = \sum_{k \in \mathbb{N}} \langle f, b_k \rangle_{L_2}\, b_k(x) \qquad \text{for all } x \in D_0 \text{ and } f \in F_0.$$

The proof: From countable to separable

$F \hookrightarrow L_2$ is a separable metric space with cont. point evaluation.

- F contains a countable dense subset F_0
- $\Big\|f - A_N(f)\Big\|_{L_2} \le \Big\|f - g\Big\|_{L_2} + \Big\|g - A_N(g)\Big\|_{L_2} + \Big\|A_N(f-g)\Big\|_{L_2}$
- $U_\delta(f) := \{g \in F \colon d_F(f,g) < \delta\}$ and $\delta > 0$ small enough
- $g \in F_0 \cap U_\delta(f)$: $\quad \|f - g\|_{L_2} < \varepsilon$ and $|f(x_i) - g(x_i)| < \varepsilon$ (!!!)
- $\Big\|f - A_N(f)\Big\|_{L_2} \le \sup_{g \in F_0} \Big\|g - A_N(g)\Big\|_{L_2} + C\varepsilon$

Hence,

$$e(A_N, F) = e(A_N, F_0) \qquad \text{for every } \underline{\text{linear}} \ A_N.$$

Downsampling

To finish the proof, we take n "good" out of $n \log n$ random points. (This was done first by [Limonova/Temlykov 2020, NSU 2020].)

That is, for some $J \subset \{1, \dots, N\}$, we consider

$$G_J := \left(\frac{b_k(x_i)}{\sqrt{\varrho(x_i)}} \right)_{i \in J,\, k \le n} \qquad \text{and} \qquad N_J(f) := \left(\frac{f(x_i)}{\sqrt{\varrho(x_i)}} \right)_{i \in J}.$$

Then, the (linear) algorithm $A_J := G_J^+ N_J$ uses only $|J|$ function values and satisfies

$$e(A_J, F) \le a_n + s_{\min}(G_J)^{-1} \sup_{f \in F_0} \|N_J(f - P_n f)\|_{\ell_2^{|J|}},$$

Downsampling II

For $J \subset \{1, \ldots, N\}$ and $f \in F$, we have $\|N_J(f)\|_{\ell_2^{|J|}} \le \|N(f)\|_{\ell_2^N}$ and hence

$$\|N_J(f - P_n(f))\|_{\ell_2^{|J|}} \le c_p \sqrt{n \log n} \left(\frac{1}{n} \sum_{k \ge n} a_k^p \right)^{1/p}.$$

It remains to find $J \subset \{1, \ldots, N\}$ with $\#J \le c_1 n$ such that

$$s_{\min}(G_J)^2 \ge c_2 n.$$

Recall that $\forall w \in \mathbb{C}^n$: $\frac{N}{2} \le \frac{\|Gw\|_2^2}{\|w\|_2^2} \le \frac{3N}{2}$ with high probability.

Downsampling III

This is based on the following fascinating result.

Weaver's theorem [Weaver '04, MSS '15, NOU '16, LT '20, NSU '20]

There exist constants $c_1, c_2, c_3 > 0$ such that, for all $u_1, \dots, u_N \in \mathbb{C}^n$ such that $\|u_i\|_2^2 \le 2n$ for all $i = 1, \dots, N$ and

$$\frac{1}{2}\|w\|_2^2 \le \frac{1}{N}\sum_{i=1}^{N} |\langle w, u_i\rangle|^2 \le \frac{3}{2}\|w\|_2^2, \qquad w \in \mathbb{C}^n,$$

there is a $J \subset \{1, \dots, m\}$ with $\#J \le c_1 n$ and

$$c_2\,\|w\|_2^2 \le \frac{1}{n}\sum_{i \in J} |\langle w, u_i\rangle|^2 \le c_3\|w\|_2^2, \qquad w \in \mathbb{C}^n.$$

(This is based on the famous solution of the Kadison-Singer problem.)

Finally...

Theorem [Krieg/U 2021]

Let $F \hookrightarrow L_2$ be a separable metric space of functions on D, such that point evaluation is continuous on F, i.e., $\{\delta_x : x \in D\} \subset F'$. Then, for every $0 < p < 2$, there is a constant $c_p > 0$, depending only on p, such that, for all $n \geq 2$, we have

$$g_N(F) \leq \sqrt{\log n} \left(\frac{1}{n} \sum_{k \geq n} a_k(F)^p \right)^{1/p}$$

for $N \geq c_p \cdot n$.

For more on the power of this 'downsampling' see Tino's talk...

My favorite example

A prominent example:

Sobolev spaces with (dominating) mixed smoothness.

Let $D = \mathbb{T}^d$ be the d-dim. torus, $\mu = \lambda$ the Lebesgue measure on $\mathbb{T}^d$, $1 \leq p \leq \infty$ and $s \in \mathbb{N}$. We define

$$\mathbf{W}_p^s = \left\{ f \in L_p(\mathbb{T}^d) \colon \|f\|_{\mathbf{W}_p^s} \leq 1 \right\},$$

where

$$\|f\|_{\mathbf{W}_p^s} := \left(\sum_{\alpha \in \mathbb{N}_0^d \colon |\alpha|_\infty \leq s} \|D^\alpha f\|_p^p \right)^{1/p}.$$

So, $f \in \mathbf{W}_p^s$ implies $D^\alpha f \in L_p$ for all $\alpha \in \mathbb{N}_0^d$ with $\max_i |\alpha_i| \leq s$.

My favorite example II

It is known that these well-studied spaces satisfy

- $g_n(\mathbf{W}^s_p) \asymp a_n(\mathbf{W}^s_p)$ for $p < 2$ and all $s > 1/p$.
- $g_n(\mathbf{W}^s_p) \geq a_n(\mathbf{W}^s_p) \asymp n^{-s} \log^{s(d-1)}(n)$ for $p \geq 2$ and $s > 0$.
- $g_n(\mathbf{W}^s_p) \lesssim n^{-s} \log^{(s+1/2)(d-1)}(n)$ for $p \geq 2$ and $s > 1/2$.

All the upper bounds are achieved by sparse grids.[Sickel, T. Ullrich, 2007]

It was the prevalent conjecture that the upper bounds are sharp.

My favorite example III

For the spaces $\mathbf{W}_p^s$ the "good" ONB is given by $\{e^{2\pi i k \cdot} \colon k \in \mathbb{Z}^d\}$, i.e. the Fourier basis. Since $\|b_k\|_\infty \lesssim 1$, we can use $\rho \equiv 1$.

Corollary [Krieg/U 2019, U 2020]

Let $x_1, \dots, x_n$ be independent and uniformly distributed in $\mathbb{T}^d$. Then, for any $s > 1/2$,

$$e(A_n, \mathbf{W}_2^s) \;\lesssim\; a_{\frac{n}{\log n}}(\mathbf{W}_2^s) \;\asymp\; n^{-s} \log^{sd}(n)$$

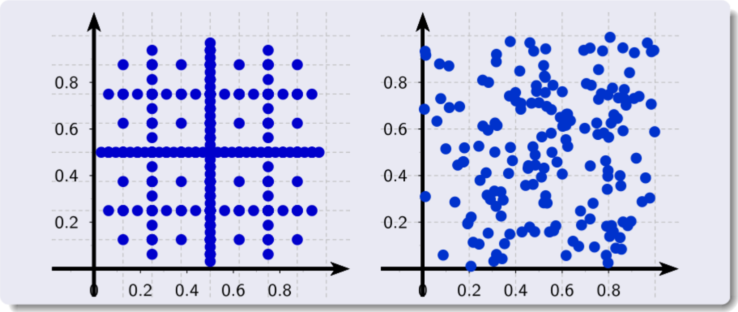
with probability at least $1 - \frac{8}{n^2}$.

Nagel/Schäfer/T. Ullrich 2020: $\quad e_n(\mathbf{W}_2^s) \;\lesssim\; n^{-s} \log^{s(d-1)+1/2}(n)$.

Sparse grids vs. random point sets

$$w.h.p.: \qquad e(A_n, \mathbf{W}_2^s) \lesssim n^{-s} \log^{sd}(n),$$

which is better than sparse grids for $d > 2s + 1$.



What are optimal points?

Good point sets

Open problems:

1. Find an explicit construction of such point sets!
2. What are necessary/sufficient conditions?

Note: Lattices don't work. Nets?

$\rightsquigarrow$ We still don't know enough about some of the easiest (general) approximation problems in high dimensions...

Special information

In the above, there's nothing special about function values, and we can do the same for **other classes on information**:

Given a class $\Lambda \subset F'$ of admissible information, let

$$a_n(F,\Lambda) := \inf_{N_n \in \Lambda^n} e(F, N_n)$$

be the n-th minimal worst-case error of linear algorithms based on optimal info from Λ.

Special info: The result

Theorem [work in progress]

Let $\Lambda \subset F'$ be such that there exist a measure ν on Λ with

$$\int_\Lambda L(f) \cdot \overline{L(g)} \, \mathrm{d}\nu(L) \; = \; \langle f, g \rangle_{L_2}$$

for all $f, g \in F$.

Then,

$$a_N(F, \Lambda) \; \leq \; \sqrt{\log n} \, \left(\frac{1}{n} \sum_{k \geq n} a_k(F)^p \right)^{1/p}$$

for $0 < p < 2$ and $N \geq c_p \cdot n$.

One obtains better bounds for more special info...

Special info: Example

Consider an **arbitrary orthonormal basis**

$$\mathcal{H} = \{h_1, h_2, \dots\} \text{ of } L_2.$$

By choosing ν to be the counting measure, we see

$$\int_\Lambda c(f) \cdot \overline{c(g)} \, \mathrm{d}\nu(c) \; = \; \sum_{i=1}^{\infty} \langle f, h_i \rangle \cdot \overline{\langle g, h_i \rangle} \; = \; \langle f, g \rangle_{L_2}.$$

$\rightsquigarrow$ In this formulation, F does not appear at all.

$\rightsquigarrow$ Your favorite L_2-basis gives almost optimal info if $(a_n) \in \ell_2$.

Special info: The algorithm

For a given class of admissible info $\Lambda \subset F'$, and given $c_1, \ldots, c_N \in \Lambda$, let

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

with

$$\varrho : \Lambda \to \mathbb{R}, \quad \varrho(c) = \frac{1}{2}\left(\frac{1}{n}\sum_{k \le n} |c(b_k)|^2 + \sum_{k>n} w_k |c(b_k)|^2\right).$$

Non-linear algorithms

One might want to consider **arbitrary algorithms**:

$$A_n(f) \,=\, \psi\Big(L_1(f),\dots,L_n(f)\Big) \,\in\, L_2$$

with some $L_1,\dots,L_n \in F'$ and a (non-linear) mapping $\psi\colon \mathbb{R}^n \to L_2$.

Gelfand width:

$$c_n(F,\Lambda) \,:=\, \inf_{\substack{\psi\colon \mathbb{R}^n\to L_2\\ L_1,\dots,L_n\in\Lambda}} \sup_{f\in F} \Big\| f - \psi\Big(L_1(f),\dots,L_n(f)\Big)\Big\|_{L_2}.$$

$$c_n(F) \,:=\, c_n(F,F')$$

Non-linear algorithms II

Let F be a unit ball of a Banach space.

Several results are known to compare these quantities:

Linear vs. non-linear: $\sup_F \left\{ \frac{a_n(F)}{c_n(F)} \right\} \asymp \sqrt{n}$

Linear vs. non-linear sampling: $\sup_F \left\{ \frac{g_n(F)}{c_n(F,\{\delta_x\})} \right\} \asymp \sqrt{n}$

Lower bound for sampling:

$$g_n(W_1^s([0,1])) \geq c_n(W_1^s([0,1]),\{\delta_x\}) \asymp 1 \text{ for } s < 1.$$

See books of Novak/Wozniakowski 08-12 (Chapter 29), Pinkus etc.

Non-linear algorithms III

Since our result implies

$$g_N(F) \;\le\; \sqrt{\log n}\,\left(\frac{1}{n}\sum_{k\ge n}\left(\sqrt{k}\,c_k(F)\right)^p\right)^{1/p}$$

for $N \ge c_p \cdot n$, we also know what happens here in the "worst case":

For F a unit ball of a Banach space, we have for $s > 1$

$$n^{-s+1/2} \lesssim \sup\Big\{g_n(F)\colon\, F \text{ with } c_n(F) \le n^{-s}\Big\} \lesssim \sqrt{\log n}\cdot n^{-s+1/2}$$

and for $s \le 1$

$$\sup\Big\{g_n(F)\colon\, F \text{ with } c_n(F) \le n^{-s}\Big\} \asymp 1$$

Final remarks

- We have a quite complete picture of the power of function values, if we only assume some decay on (a_n) or (c_n).
- What about other (general) assumptions? (See e.g. Jan's talk)
- Is the $\sqrt{\log(n)}$-factor needed?
- Can non-linear algorithms do "better"?
- Again: What are good point sets?

Thank you!

History: The simplex

$$B_1^m = \left\{ x \in \mathbb{R}^m \;\middle|\; \sum_{j=1}^m |x_j| \le 1 \right\}.$$

Theorem (Kashin, Garnaev, Gluskin)

Consider the recovery of vectors from B_1^m in the Euclidean norm with Gaussian information. Then

$$\mathbb{E}\left[e(B_1^m, N_n)\right] \asymp c_n(B_1^m) \asymp \min\left\{1, \sqrt{\frac{\log(1+\frac{m}{n})}{n}}\right\}.$$

An analogous estimate holds with high probability.

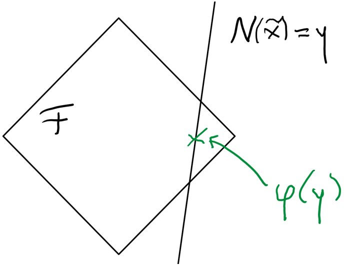
Although most of the information mappings yield optimal information, not a single example is known explicitly.

The bound is achieved by the algorithm

$$A_n(x) = \varphi(N_n(x))$$

with the nonlinear mapping

$$\varphi(y) = \operatorname*{argmin}_{\tilde{x}\in\mathbb{R}^m:\ N_n(\tilde{x})=y} \|\tilde{x}\|_1.$$



That is, we have

$$\mathbb{E}[e(A_n, B_1^m)] \asymp \min\left\{1, \sqrt{\frac{\log(1+\frac{m}{n})}{n}}\right\}.$$

It is known that linear algorithms are much worse. We have

$$a_n(B_1^m) = \left(\frac{m-n}{m}\right)^{1/2}.$$

Why mixed smoothness?

Spaces with mixed smoothness are of interest (for numerics) because they ...

- are tensor products of univariate spaces.
- correspond to several concepts of "uniform distribution theory".
- reflect the independence of parameters in high-dimensional models, like medical data, physical measurements etc.
- are proven to be important for the electronic Schrödinger equation. [Yserentant, 2005]