Learning a Function from Noisy Samples at a Finite Sparse Set of Points

Andreas Hofinger [a,1,*], Friedrich Pillichshammer [b,2]

[a] *Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria*

[b] *University of Linz, Institute for Financial Mathematics, Altenbergerstraße 69, A-4040 Linz, Austria*

Abstract

In learning theory the goal is to reconstruct a function defined on some (typically high-dimensional) domain Ω, when only noisy values of this function at a sparse, discrete subset $\omega \subset \Omega$ are available.

In this work we use Koksma-Hlawka type estimates to obtain deterministic bounds on the so-called *generalization error*. The resulting estimates show, that the generalization error tends to zero, when the noise in the measurements tends to zero, and the number of sampling points tends to infinity sufficiently fast. None of the obtained rates does depend on the dimension of the sampling space.

Key words: Sampling theory, Learning theory, Regularization, quasi-Monte Carlo methods.
1991 MSC: 94A20, 68T05, 65C05, 47A52

1 Introduction

In the problem of learning a function f, the aim is to generalize knowledge available on a discrete set $\omega = \{x_1, \ldots, x_N\} \subset \Omega$ onto the whole domain Ω. Several attempts have been made to determine bounds for the resulting *generalization error*, the aim of this work is to develop a framework that allows to bound this error under simple and verifiable assumptions.

* Corresponding author
Email addresses: andreas.hofinger@oeaw.ac.at (Andreas Hofinger), friedrich.pillichshammer@jku.at (Friedrich Pillichshammer).

[1] Supported by the Austrian Research Fund (FWF), Project SFB F 013/08.
[2] Supported by the Austrian Research Fund (FWF), Project P17022-N12.

The problem of interest can be stated as

$$\begin{gathered}\text{Given discrete values } f^{\delta}(x_i),\ x_i \in \omega,\ i = 1, \ldots, N, \\ \text{find an approximation } f^{\delta}_{\omega} \text{ to } f \text{ on the whole domain } \Omega.\end{gathered} \tag{1}$$

For the given point measurements $f^{\delta}(x_i)$ we assume

$$f^{\delta}(x_i) = f(x_i) + \delta_i, \quad x_i \in \omega = \{x_1, \ldots, x_N\},$$

where δ_i represents noise. To keep our approach as general as possible, we will not make any assumptions on the nature of the perturbations δ_i (such as independence or boundedness), besides the requirement that

$$\frac{1}{N} \sum_{i=1}^{N} \left(f^{\delta}(x_i) - f(x_i) \right)^2 \leq \delta^2 < \infty.$$

This is a natural assumption and fulfilled in typical applications (see below).

To obtain our results, smoothness assumptions on the function f are necessary; these will be stated via norms in *Sobolev spaces* (cf. [1]). In the following, the Sobolev space $H^s(\Omega)$ is defined as

$$H^s(\Omega) = \left\{ h \in L_2(\Omega) \mid \|h\|_{H^s(\Omega)} < \infty \right\},$$

where the corresponding norm is given by

$$\|h\|_{H^s(\Omega)} = \left(\sum_{0 \leq |\mathfrak{u}| \leq s} \int_{\Omega} \left(\frac{\partial^{|\mathfrak{u}|}}{\partial x_{\mathfrak{u}}} h(x) \right)^2 \mathrm{d}x \right)^{1/2}, \tag{2}$$

i. e., the sum of the $L_2(\Omega)$-norm of all weak derivatives of h up to order s. It should be mentioned that this definition can be extended to non-integers s as well, for details we refer to [1, Chapter 7].

The assumptions we will need to derive bounds on the generalization error are extremely simple, and thus also interpretable and can be verified in practice; nevertheless, fundamental results from the theory of quasi-Monte Carlo integration are necessary to obtain these estimates. To abbreviate notation we introduce the $\ell_2(\omega)$-norm on the discrete set $\omega = \{x_1, \ldots, x_N\}$ as

$$\|h\|^2_{\ell_2(\omega)} := \frac{1}{N} \sum_{i=1}^{N} h(x_i)^2.$$

The necessary assumptions on the data $\omega = \{x_1, \ldots, x_N\}$, $f^{\delta}(x_i)$, and the constructed approximation f^{δ}_{ω} are then given as follows.

Assumption 1 (with parameter s) *The noisy measurements $f^\delta(x_i)$, taken at points $\omega = \{x_1, \dots, x_N\}$ are a discrete approximation to $f \in L_2(\Omega)$ and have $\ell_2(\omega)$ noise level*

$$\left\| f - f^\delta \right\|_{\ell_2(\omega)}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(f(x_i) - f^\delta(x_i) \right)^2 \le \delta^2 . \tag{3a}$$

The approximation $f_\omega^\delta \in L_2(\Omega)$ constructed from $f^\delta(x_i)$ satisfies

$$\left\| f - f_\omega^\delta \right\|_{\ell_2(\omega)}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(f(x_i) - f_\omega^\delta(x_i) \right)^2 \le C_1 \delta^2 . \tag{3b}$$

The distance of the approximation f_ω^δ to the function f, measured in the Sobolev space $H^s(\Omega)$ is bounded, i. e.,

$$\left\| f - f_\omega^\delta \right\|_{H^s(\Omega)} \le C_2 . \tag{3c}$$

The constants C_1 and C_2 are both independent of ω.

Under these simple assumptions we are able to derive our main results, Theorems 8 and 12 below: when the noise level δ tends to zero, and N increases sufficiently fast, then also the error measured on the whole space $L_2(\Omega)$ tends to 0. In the following we give a short discussion of these assumptions.

Remark 2 Observe that condition (3a) is fulfilled in very general cases, for instance pointwise bounded errors as in [2,3] are allowed. But in contrast to their work, here also the important case of pointwise Gaussian measurement errors is permitted, more general, any independent, identically distributed (i.i.d.) perturbation with bounded variance satisfies (3a).

The constant C_1 in condition (3b) will always be strictly greater than 1, since it is in general not possible to satisfy (3b) and (3c) simultaneously, when $C_1 = 1$.

Finally we would like to mention that the smoothness assumption (3c) required on f_ω^δ can be verified easily, it boils down to the requirement that:

- The true solution f carries some smoothness.
- The constructed approximation f_ω^δ is smooth as well (but not necessarily with the same index s).
- The procedure to generate f_ω^δ from measurements in ω keeps f_ω^δ smooth. (For the procedures we consider in Section 4, condition (3c) is satisfied with $C_2 \le 2\|f\|_{H^s(\Omega)}$.)

Observe that no assumptions are made on how the function f_ω^δ are obtained; two possibilities to generate functions f_ω^δ with the desired properties are described in

Section 4, but also other methods will work as well, when the regularization parameters are chosen appropriately. In earlier works different assumptions were necessary, many of them difficult to check.

For example, the estimates in [4] require bounds on the Vapnik-Chervonenkis-dimension of certain sets, the results in [5] needed the concept of covering numbers and pseudo-dimensions; also in [2] the concept of covering numbers is used (cf. e. g., [2, Theorem B]). A conceptually very different approach was taken in [3] were the focus was on reproducing kernel Hilbert spaces; using such spaces one can represent linear operators as infinite matrices, to bound the generalization error, estimates on infimal and supremal singular values of such infinite matrices with random entries were necessary.

In contrast, the assumptions in our setup have simple interpretations and are not restricted to a particular learning scheme (like e. g., regularization networks). Any learning method that satisfies Assumptions 1 is allowed. It should also be mentioned, that in the case of neural networks the assumptions reduce to equivalent smoothness requirements on the activation function Φ in (10) (cf. Section 4).

The outline of this paper is as follows: Section 2 is devoted to the derivation of a Koksma-Hlawka type estimate for bounding the generalization error. These results are applied in Section 3 to obtain our main result, Theorem 8: a dimension independent bound on the generalization error. In the second part of this section, results from the theory of Sobolev spaces are used, to deduce the same convergence rate under a weaker growth condition on the number N of samples. In Section 4 we present two possibilities to generate functions f_ω^δ with the properties described in Assumptions 1. Finally, in Section 5 we give an outlook on possibilities for future work and further improvements.

2 A Koksma-Hlawka type estimate for bounding the Generalization Error

In this section we derive a Koksma-Hlawka type bound on the generalization error. Based on Assumptions 1, this estimate can be used to obtain a probabilistic bound on the $L_2(\Omega)$-norm of the error in Section 3.

In the sequel, for simplicity we shall restrict ourselves to the case $\Omega = [0,1]^d$, the d-dimensional unit cube. (Since we consider the case of weighted Sobolev-spaces, the analysis can easily be extended to more general domains.) It is clear that a good approximation to f in $L_2(\Omega)$ can only obtained, if the sampled points cover the whole cube, nevertheless, to allow more general cases we introduce a probability density function $\rho(x)$ on the d-dimensional unit cube $[0,1]^d$, and assume that the points x_i are drawn from this distribution. Consequently the resulting estimates for the error will be given in terms of weighted $L_{2,\rho}(\Omega)$-norms.

Observe that the error in the $L_{2,\rho}(\Omega)$-norm can be split into one part that measures the approximation quality on the discrete set $\omega = \{x_1, \ldots, x_N\}$, and another part that directly measures the generalization error:

$$
\begin{aligned}
&\left\| f - f_\omega^\delta \right\|^2_{L_{2,\rho}(\Omega)} \\
&\leq \left\| f - f_\omega^\delta \right\|^2_{\ell_2(\omega)} + \left| \left\| f - f_\omega^\delta \right\|^2_{L_{2,\rho}(\Omega)} - \left\| f - f_\omega^\delta \right\|^2_{\ell_2(\omega)} \right| \qquad (4)\\
&= \left\| f - f_\omega^\delta \right\|^2_{\ell_2(\omega)} + \left| \int_{[0,1]^d} (f(x) - f_\omega^\delta(x))^2 \rho(x)\, \mathrm{d}x - \frac{1}{N} \sum_{i=1}^N (f(x_i) - f_\omega^\delta(x_i))^2 \right|.
\end{aligned}
$$

According to Assumptions 1 we may assume that the first term in (4) is bounded by $\|f - f_\omega^\delta\|^2_{\ell_2(\omega)} \leq C\delta^2$. This is for instance the case when f_ω^δ is generated via the greedy algorithm discussed in Section 4; the goal of this section is to obtain bounds for the second term. This part can be interpreted as the error obtained by a quasi-Monte Carlo integration rule, where the function to be integrated is given by $(f - f_\omega^\delta)^2$ (for an introduction in quasi-Monte Carlo rules see, for example, [6]). To estimate this term we will use a Koksma-Hlawka type estimate, which separates the bound into properties of the point set ω and smoothness properties of the function $(f - f_\omega^\delta)^2$.

We introduce some notation: let $\mathcal{D}$ denote the index set $\mathcal{D} = \{1, \ldots, d\}$. For $\mathfrak{u} \subseteq \mathcal{D}$ let $|\mathfrak{u}|$ denote the cardinality of $\mathfrak{u}$ and for a vector $x \in I^d := [0,1]^d$ let $x_\mathfrak{u}$ denote the vector from $I^{|\mathfrak{u}|}$ containing all components of x whose indices are in $\mathfrak{u}$. Further let $\mathrm{d}x_\mathfrak{u} = \prod_{j \in \mathfrak{u}} \mathrm{d}x_j$ and let $(x_\mathfrak{u}, 1)$ be the vector x from I^d with all components whose indices are not in $\mathfrak{u}$ replaced by 1.

The integration error depends on smoothness assumptions of the function to be integrated. For this sake, let the $2,d$-variation of a function h be defined as

$$
\|h\|_{2,d} := \left(\sum_{\mathfrak{u} \subseteq \mathcal{D}} \int_{[0,1]^{|\mathfrak{u}|}} \left| \frac{\partial^{|\mathfrak{u}|}}{\partial t_\mathfrak{u}} h(t_\mathfrak{u}, 1) \right|^2 \mathrm{d}t_\mathfrak{u} \right)^{1/2}, \qquad (5)
$$

and the corresponding function space be denoted by

$$
\mathcal{F}_{2,d} := \{ h \in L_2([0,1]^d) : \|h\|_{2,d} < \infty \}.
$$

The integration error can now be bounded by the variation of the function to be integrated, and the discrepancy of the points used for this integration, via the following Koksma-Hlawka type inequality. To abbreviate notation we write $\mathrm{d}_\rho(x)$ instead of $\rho(x)\,\mathrm{d}x$ from now on.

Proposition 3 *Let $h \in \mathcal{F}_{2,d}$ and let $x_1, \ldots, x_N$ be N points in $[0,1)^d$. Then we have*

$$
\left| \int_{[0,1]^d} h(x)\, \mathrm{d}_\rho(x) - \frac{1}{N} \sum_{n=1}^N h(x_n) \right| \leq \|h\|_{2,d} D_{2,\rho,N}(x_1, \ldots, x_N),
$$

where

$$D_{2,\rho,N}(x_1,\dots,x_N) := \left(\sum_{\mathfrak{u} \subseteq \mathcal{D}} \int_{[0,1]^{|\mathfrak{u}|}} \mathrm{disc}_\rho(t_\mathfrak{u},1)^2 \,\mathrm{d}t_\mathfrak{u} \right)^{1/2}$$

and

$$\mathrm{disc}_\rho(x) := \int_{[0,1]^d} \chi_{[0,x)}(t)\,\mathrm{d}\rho(t) - \frac{1}{N}\sum_{i=1}^N \chi_{[0,x)}(x_i).$$

Here $\chi_{[0,x)}$ denotes the characteristic function of the interval $[0,x)$. We shall call $D_{2,\rho,N}$ the $L_{2,\rho}$-discrepancy of the points $x_1,\dots,x_N$.

PROOF. Let $h \in \mathcal{F}_{2,d}$ and let $x_1,\dots,x_N$ be N points in $[0,1)^d$. Then it follows from [7, Eq. (12)] that

$$\int_{[0,1]^d} h(x)\,\mathrm{d}\rho(x) - \frac{1}{N}\sum_{n=1}^N h(x_n) = \sum_{\mathfrak{u}\subseteq\mathcal{D}} (-1)^{|\mathfrak{u}|} \int_{[0,1]^{|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}}{\partial t_\mathfrak{u}} h(t_\mathfrak{u},1)\mathrm{disc}_\rho(t_\mathfrak{u},1)\,\mathrm{d}t_\mathfrak{u}.$$

As in [7] an application of the Cauchy-Schwarz inequality yields the result. □

We obtain the following probabilistic results for the $L_{2,\rho}$-discrepancy of random points in the unit cube.

Proposition 4 *Assume that $x_1,\dots,x_N$ are N independent random variables with density ρ on $[0,1]^d$. Denote by $\mathbb{E}$ (resp. $\mathbb{P}$) the expectation (resp. the probability) with respect to the density ρ on $[0,1]^d$.*

(1) Then the mean square $L_{2,\rho}$-discrepancy is given by

$$\mathbb{E}[D^2_{2,\rho,N}(x_1,\dots,x_N)] = \frac{\Lambda_{\rho,d}}{N},$$

where

$$\Lambda_{\rho,d} := \sum_{\mathfrak{u}\subseteq\mathcal{D}} \Bigg[\int_{[0,1]^{|\mathfrak{u}|}} \int_{[0,1]^d} \chi_{[0,(t_\mathfrak{u},1))}(x)\,\mathrm{d}\rho(x)\,\mathrm{d}t_\mathfrak{u} - \int_{[0,1]^{|\mathfrak{u}|}} \left(\int_{[0,1]^d} \chi_{[0,(t_\mathfrak{u},1))}(x)\,\mathrm{d}\rho(x) \right)^2 \mathrm{d}t_\mathfrak{u} \Bigg].$$

(2) For any $c > 1$ we have

$$\mathbb{P}\left[D_{2,\rho,N}(x_1,\dots,x_N) < c\sqrt{\Lambda_{\rho,d}/N} \right] \geq 1 - \frac{1}{c^2}.$$

Remark 5 If the density ρ is of product form, i.e., for $x = (x_1, \dots, x_d)$, $\rho(x) = \prod_{k=1}^{d} \rho_k(x_k)$ with probability densities ρ_k on $[0,1]$, then we obtain

$$\Lambda_{\rho,d} = \prod_{k=1}^{d}(1+\alpha_k) - \prod_{k=1}^{d}(1+\beta_k),$$

where

$$\alpha_k = 1 - \int_0^1 t\rho_k(t)\,\mathrm{d}t \qquad \text{and} \qquad \beta_k = \int_0^1 \left(\int_0^t \rho_k(x)\,\mathrm{d}x\right)^2 \mathrm{d}t.$$

In particular, if $\rho \equiv 1$, i.e., the points are uniformly distributed on the unit cube, then $\Lambda_{\rho,d} = \left(\frac{3}{2}\right)^d - \left(\frac{4}{3}\right)^d$.

PROOF. From the linearity of expectation we obtain

$$\mathbb{E}[D^2_{2,\rho,N}(x_1,\dots,x_N)] = \sum_{\mathfrak{u}\subseteq\mathcal{D}} \mathbb{E}[\Delta(\mathfrak{u})],$$

where $\Delta(\mathfrak{u})$ is defined as

$$\Delta(\mathfrak{u}) := \int_{[0,1]^{|\mathfrak{u}|}} \mathrm{disc}_\rho(t_\mathfrak{u},1)^2\,\mathrm{d}t_\mathfrak{u}.$$

Using the binomial formula, the square of the function $\mathrm{disc}_\rho(t_\mathfrak{u},1)$ can be written as

$$\begin{aligned}
\mathrm{disc}_\rho(t_\mathfrak{u},1)^2 &= \left(\int_{[0,1]^d} \chi_{[0,(t_\mathfrak{u},1))}(x)\,\mathrm{d}\rho(x) - \frac{1}{N}\sum_{i=1}^{N}\chi_{[0,(t_\mathfrak{u},1))}(x_i)\right)^2 \\
&= \left(\int_{[0,1]^d} \chi_{[0,(t_\mathfrak{u},1))}(x)\,\mathrm{d}\rho(x)\right)^2 \\
&\quad - \frac{2}{N}\sum_{i=1}^{N}\int_{[0,1]^d} \chi_{[0,(t_\mathfrak{u},1))}(x)\,\mathrm{d}\rho(x)\chi_{[0,(t_\mathfrak{u},1))}(x_i) \\
&\quad + \frac{1}{N^2}\sum_{i,j=1}^{N}\chi_{[0,(t_\mathfrak{u},1))}(x_i)\chi_{[0,(t_\mathfrak{u},1))}(x_j).
\end{aligned}$$

Therefore for the expectation of $\Delta(\mathfrak{u})$ we obtain

$$\begin{aligned}
&\mathbb{E}[\Delta(\mathfrak{u})] = \\
&= \int_{[0,1]^{|\mathfrak{u}|}} \left(\int_{[0,1]^d} \chi_{[0,(t_\mathfrak{u},1))}(x)\,\mathrm{d}\rho(x)\right)^2 \mathrm{d}t_\mathfrak{u} \\
&\quad - \frac{2}{N}\sum_{i=1}^{N}\int_{[0,1]^d}\int_{[0,1]^{|\mathfrak{u}|}}\int_{[0,1]^d} \chi_{[0,(t_\mathfrak{u},1))}(x)\,\mathrm{d}\rho(x)\chi_{[0,(t_\mathfrak{u},1))}(x_i)\,\mathrm{d}t_\mathfrak{u}\,\mathrm{d}\rho(x_i)
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{N^2}\sum_{i,j=1}^{N}\mathbb{E}[\chi_{[0,(t_{\mathfrak{u}},1))}(x_i)\chi_{[0,(t_{\mathfrak{u}},1))}(x_j)]\\
=&-\int_{[0,1]^{|\mathfrak{u}|}}\left(\int_{[0,1]^d}\chi_{[0,(t_{\mathfrak{u}},1))}(x)\,\mathrm{d}_\rho(x)\right)^2\mathrm{d}t_{\mathfrak{u}}\\
&+\frac{1}{N^2}\sum_{i=1}^{N}\int_{[0,1]^d}\int_{[0,1]^{|\mathfrak{u}|}}\chi_{[0,(t_{\mathfrak{u}},1))}(x_i)\,\mathrm{d}t_{\mathfrak{u}}\,\mathrm{d}_\rho(x_i)\\
&+\frac{1}{N^2}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\int_{[0,1]^d}\int_{[0,1]^d}\int_{[0,1]^{|\mathfrak{u}|}}\chi_{[0,(t_{\mathfrak{u}},1))}(x_i)\chi_{[0,(t_{\mathfrak{u}},1))}(x_j)\,\mathrm{d}t_{\mathfrak{u}}\,\mathrm{d}_\rho(x_i)\,\mathrm{d}_\rho(x_j)\\
=&-\int_{[0,1]^{|\mathfrak{u}|}}\left(\int_{[0,1]^d}\chi_{[0,(t_{\mathfrak{u}},1))}(x)\,\mathrm{d}_\rho(x)\right)^2\mathrm{d}t_{\mathfrak{u}}\\
&+\frac{1}{N}\int_{[0,1]^d}\int_{[0,1]^{|\mathfrak{u}|}}\chi_{[0,(t_{\mathfrak{u}},1))}(x)\,\mathrm{d}t_{\mathfrak{u}}\,\mathrm{d}_\rho(x)\\
&+\frac{N^2-N}{N^2}\int_{[0,1]^{|\mathfrak{u}|}}\left(\int_{[0,1]^d}\chi_{[0,(t_{\mathfrak{u}},1))}(x)\,\mathrm{d}_\rho(x)\right)^2\mathrm{d}t_{\mathfrak{u}}\\
=&\frac{1}{N}\Bigg[\int_{[0,1]^d}\int_{[0,1]^{|\mathfrak{u}|}}\chi_{[0,(t_{\mathfrak{u}},1))}(x)\,\mathrm{d}t_{\mathfrak{u}}\,\mathrm{d}_\rho(x)\\
&\qquad-\int_{[0,1]^{|\mathfrak{u}|}}\left(\int_{[0,1]^d}\chi_{[0,(t_{\mathfrak{u}},1))}(x)\,\mathrm{d}_\rho(x)\right)^2\mathrm{d}t_{\mathfrak{u}}\Bigg].
\end{aligned}
$$

The first result follows. The second result follows from the first one together with the Markov inequality (see e. g., [8]). □

Now we have all tools that we need for bounding the generalization error. This will be done in the subsequent section.

3 Bounding the Generalization Error

Using the results of the previous section we can now derive a bound on the generalization error. First of all we will connect the $2,d$-variation with Sobolev-norms as defined in (2). The reason for using these Sobolev norms instead of the $2,d$-norm directly, is that for the former, *interpolation inequalities* ([1,9,10]) are available. These inequalities will be necessary to obtain the results of Theorem 8 under the weaker assumptions of Theorem 12.

Proposition 6 *Let $h \in H^s(\Omega)$ with $\Omega = [0,1]^d$ and $s \geq d$. Then there exists $C \in \mathbb{R}$ with*

$$\|h^2\|_{2,d} \leq C\|h\|^2_{H^s(\Omega)}. \tag{6}$$

PROOF. The proof follows from definitions (2) and (5) using the Sobolev embedding theorem (see e. g. [1]). Recall that this theorem states, that the $W^{j,q}(\Omega_k)$-norm of h can be bounded by its $W^{j+m,p}(\Omega)$-norm, whenever $p \leq q \leq kp/(d-mp)$ and $d > mp$. (Here by Ω_k we denote a k-dimensional subset of Ω.) Since we are only interested in the index s necessary to bound the variation, we do not have to distinguish with respect to which variable derivatives are built, only the number of derivatives is important. Therefore we will use the abbreviation $\frac{\partial^{|\mathfrak{u}|}}{x_{\mathfrak{u}}} h =: h_{|\mathfrak{u}|}$. Using Leibniz' identity (the product rule for higher order derivatives) we obtain

$$
\begin{aligned}
\|h^2\|_{2,d}^2 &= (h(1)^2)^2 + \sum_{k=1}^{d} \int_{\Omega_k} c_k \left((h^2)_k \right)^2 \\
&= h(1)^4 + \sum_{k=1}^{d} \int_{\Omega_k} c_k \left(\sum_{i=0}^{k} \binom{k}{i} h_{k-i} h_i \right)^2 \\
&= h(1)^4 + \sum_{k=1}^{d} \int_{\Omega_k} \sum_{i,j=0}^{k} c_{k,i,j} h_{k-i} h_i h_{k-j} h_j ,
\end{aligned}
$$

with appropriate constants c_k and $c_{k,i,j}$. To ensure that the integral is bounded, it is necessary that $h_{k-i}h_i$ is in $L^2(\Omega_k)$ for all appearing combinations of i and k.

Suppose that $h \in H^s(\Omega) \equiv W^{s,2}(\Omega)$ with $s \geq d$. Via the embedding theorem this implies (if $d > 1$), $h \in W^{s-1/2,2}(\Omega_{d-1})$, and therefore of course also $h \in W^{s-1,2}(\Omega_{d-1})$; inductively this gives $h \in W^{k,2}(\Omega_k)$.

Using the assumption $h \in W^{k,2}(\Omega_k)$ we will now deduce that for every i, $h_{k-i}h_i \in L^2(\Omega_k)$. Clearly the first multiplicand is in $W^{i,2}(\Omega_k)$, while the second lies in $W^{k-i,2}(\Omega_k)$. Suppose that $(k-i) > i$, then from $h_{k-i} \in W^{i,2}(\Omega_k)$ we obtain $h_{k-i} \in L^{2k/(k-2i)}(\Omega_k) \subset L^2(\Omega_k)$, while at the same time $h_i \in L^\infty(\Omega_k)$; together this implies $h_{k-i}h_i \in L^2(\Omega_k)$. Analogous reasoning applies when $i > (k-i)$. For the case $(k-i) = i = k/2$ we use the fact that $h_{k/2} \in W^{k/2,2}(\Omega_k) \subset W^{k/4,4}(\Omega_k) \subset L^4(\Omega_k)$; since now $h_{k/2} \in L^4(\Omega_k)$ we have $h_{k/2}^2 \in L^2(\Omega_k)$.

The integral above can now be estimated using the Cauchy-Schwarz inequality and applying the obtained results to the appearing terms. For instance for terms with $(k-i) > i$ we would use the estimate

$$
\left(\int_{\Omega_k} (h_{k-i}h_i)^2 \right)^{1/2} \overset{e.g.}{\leq} \|h_{k-i}\|_{L^2(\Omega_k)} \|h_i\|_{L^\infty(\Omega_k)} \leq c\|h\|_{W^{k,2}(\Omega_k)}^2 \leq \tilde{c}\|h\|_{W^{s,2}(\Omega)}^2 .
$$

Finally we consider the appearing term $h(1)^4$. For this one we use the fact that $h \in H^s(\Omega)$ is continuous on $\bar{\Omega}$ whenever $s > d/2$. Since we assumed $s \geq d$ there is some c for which we may estimate $h(1)^4 \leq \|h\|_{C(\bar{\Omega})}^4 \leq c\|h\|_{W^{s,2}(\Omega)}^4$. Altogether this proves the desired estimate. $\square$

Remark 7 Note that the $2,d$-variation above contains the term $\int_\Omega (h_0 h_d)^2$. Even if $h_0 \in L^\infty(\Omega)$ we must require $h_d \in L^2(\Omega)$ or equivalently $h \in H^d(\Omega) = W^{d,2}(\Omega)$ to obtain a bound on the integral. Therefore the bound (6) is optimal with respect to the indices used, and can not be obtained for any $s < d$ or $p < 2$.

Although the smoothness required in Proposition 6 may seem large, recall that we already need $f \in H^s(\Omega)$ with $s > \frac{d}{2}$ to have continuous f, which is of course vital if we want to consider point measurements. Thus the smoothness requirement in Proposition 6 is only slightly stronger than the necessary smoothness assumptions for having a well-defined sampling problem.

Combining Proposition 6 with the results of the previous section we obtain the main theorem: starting with measurements on a discrete sparse set ω we obtain—under simple smoothness assumptions—convergence of f_ω^δ to f on the whole domain Ω. In particular we have a deterministic bound, as well as a bound on the expected value of the error, measured in the $L_{2,\rho}(\Omega)$-norm.

Theorem 8 *Let $\Omega = [0,1]^d$ and let Assumptions 1 with $s \geq d$ be satisfied. Then for $\delta \to 0$ the error satisfies*

$$\left\| f - f_\omega^\delta \right\|_{L_{2,\rho}(\Omega)}^2 \leq \widetilde{C} D_{2,\rho,N}(x_1, \ldots, x_N) \left\| f - f_\omega^\delta \right\|_{H^s(\Omega)}^2 + C\delta^2,$$

where C and $\widetilde{C}$ are absolute positive constants. Suppose furthermore that $x_1, \ldots, x_N$ are N independent random variables with density ρ on $[0,1]^d$, and that $N \geq \delta^{-4}$. Then for $\delta \to 0$ the expected value of the error satisfies

$$\mathbb{E}\left[\left\| f - f_\omega^\delta \right\|_{L_{2,\rho}(\Omega)}^2 \right] = O\left(\delta^2\right).$$

PROOF. For any point set $\omega = \{x_1, \ldots, x_N\} \subset \Omega$ we obtain from (4), and Propositions 3 and 6 the estimate

$$\left\| f - f_\omega^\delta \right\|_{L_{2,\rho}(\Omega)}^2 \leq \widetilde{C} D_{2,\rho,N}(x_1, \ldots, x_N) \left\| f - f_\omega^\delta \right\|_{H^s(\Omega)}^2 + C\delta^2.$$

Since $\|f - f_\omega^\delta\|_{H^s(\Omega)}$ remains bounded independent of the choice of ω, and since $\mathbb{E}[D_{2,\rho,N}(x_1, \ldots, x_N)] \leq \mathbb{E}\left[D_{2,\rho,N}^2(x_1, \ldots, x_N)\right]^{1/2}$ due to Jensen's inequality, the desired result follows from Proposition 4. □

In principle it would be possible to derive this result using the $2,d$-norm only. Nevertheless to obtain the refined estimate of Theorem 12, so-called interpolation inequalities are necessary, which are not available for the $2,d$-norm. Using additional smoothness assumptions the same convergence rate can be obtained under a weaker growth-condition on N as well. The reason for this is that convergence in $L_{2,\rho}(\Omega)$ and boundedness in higher spaces $H^r(\Omega)$ imply also convergence in intermediate spaces. Therefore in this case the term $\|f - f_\omega^\delta\|_{H^s(\Omega)}$ in Theorem 8 will not only remained bounded as stated above, but even tend to zero. How fast this convergence will be is determined by interpolation inequalities, in the following stated for weighted spaces.

Lemma 9 (Interpolation inequality) *Let $h \in H^r(\Omega)$ with $r \geq s \geq 0$. Furthermore let the density $\rho(x)$ be non-zero for all $x \in \Omega$. Then there exists some $C \in \mathbb{R}$ such that the $H^s(\Omega)$-norm of h is bounded by*

$$\|h\|_{H^s(\Omega)} \leq C\|h\|_{L_{2,\rho}(\Omega)}^{\frac{r-s}{r}} \|h\|_{H^r(\Omega)}^{\frac{s}{r}}.$$

PROOF. Since $\rho(x)$ is non-zero for all $x \in \Omega$ and since Ω is compact there exists $C_1 \in \mathbb{R}$ with

$$\|h\|_{L_2(\Omega)} \leq C_1 \|h\|_{L_{2,\rho}(\Omega)}.$$

We can now apply the interpolation inequality (see e. g., [1,9,10]) to give a bound for $\|h\|_{H^s(\Omega)}$ as

$$\|h\|_{H^s(\Omega)} \leq C_2 \|h\|_{L_2(\Omega)}^{\frac{r-s}{r}} \|h\|_{H^r(\Omega)}^{\frac{s}{r}}$$

with some C_2 depending on r, s and properties of the domain $\Omega \subset \mathbb{R}^n$. The result now follows by setting $C := C_2 C_1^{(r-s)/r}$. $\square$

Using this interpolation inequality we may rewrite Theorem 8 as follows.

Corollary 10 *Let $\Omega = [0,1]^d$ and Assumptions 1 be fulfilled with parameter r where $r \geq s \geq d$. Furthermore let the density $\rho(x)$ be non-zero for all $x \in \Omega$. Then the error satisfies the estimate*

$$\left\| f - f_\omega^\delta \right\|_{L_{2,\rho}(\Omega)}^2 \leq \widetilde{C} D_{2,\rho,N}(x_1,\ldots,x_N) \left\| f - f_\omega^\delta \right\|_{L_{2,\rho}(\Omega)}^{2\frac{r-s}{r}} + C\delta^2 \tag{7}$$

Observe that the error-term now appears on both sides of the inequality, but with different exponents. In the following we will apply a bootstrapping argument to get rid of the $L_{2,\rho}(\Omega)$-norm appearing on the right hand side of (7). To shorten the result we use some abbreviations in the following lemma, the final result in original notation is given in Theorem 12.

Lemma 11 *Let A, D, C and $\alpha \in \mathbb{R}$ be positive and $\alpha < 2$. Then the estimate*

$$A^2 \leq DA^\alpha + C \tag{8}$$

implies the bound

$$A^2 \leq D^{\frac{2}{2-\alpha}} + \frac{2}{2-\alpha} C. \tag{9}$$

PROOF. The proof follows using the (weighted) inequality of arithmetic and geometric means. Recall that this inequality states that for all $\sigma, \tau \geq 0$ and $0 < r < 1$ we have the bound

$$\sigma^r \tau^{1-r} \leq r\sigma + (1-r)\tau.$$

Since the left hand side of (8) involves the square of A, we use this estimate for the setting $\sigma := A^2$. To estimate the term DA^α we have to define τ accordingly. As turns out with the choice $\tau := D^{2/(2-\alpha)}$ we may write

$$DA^\alpha = D^{\frac{2}{2-\alpha}\left(1-\frac{\alpha}{2}\right)} A^{2\frac{\alpha}{2}} \leq \left(1-\frac{\alpha}{2}\right) D^{\frac{2}{2-\alpha}} + \frac{\alpha}{2} A^2.$$

(Since we assumed $\alpha < 2$, $r := 1-\alpha/2$ satisfies $0 < r < 1$). Combining this estimate with (8) we have

$$A^2 \leq \left(1-\frac{\alpha}{2}\right) D^{\frac{2}{2-\alpha}} + \frac{\alpha}{2} A^2 + C,$$

and since $\alpha < 2$ also

$$A^2 \leq D^{\frac{2}{2-\alpha}} + \frac{2}{2-\alpha} C,$$

which is the desired result. $\square$

Using the bootstrapping argument of Lemma 11 we are now able to show that the same convergence rate as in Theorem 8 can be obtained under a weaker growth condition on N also. Instead of $N \geq \delta^{-4}$ we now only have to require $N \geq \delta^{-4\frac{s}{r}} \geq \delta^{-2}$.

Theorem 12 *Let $\Omega = [0,1]^d$, let Assumptions 1 with r, $r \geq s \geq d$ be fulfilled and let $N \geq \delta^{-4\frac{s}{r}}$ for $s \leq r < 2s$, and $N \geq \delta^{-2}$ for $r \geq 2s$. Furthermore let the density $\rho(x)$ be non-zero for all $x \in \Omega$. Then for $\delta \to 0$ the expected value of the error satisfies the convergence rate*

$$\mathbb{E}\left[\left\|f - f_\omega^\delta\right\|^2_{L_{2,\rho}(\Omega)}\right] = O\left(\delta^2\right).$$

PROOF. After replacing the parameters in Lemma 11 by their counterparts in estimate (7) we find that the only non-deterministic entry on the right-hand side of (9) is the discrepancy. We set $\alpha = 2(r-s)/r$, where we only use smoothness up to $r \leq 2s$ to have $0 \leq \alpha \leq 1$. With this α we may apply Jensen's inequality and have

$$\mathbb{E}[D_{2,\rho,N}(x_1,\ldots,x_N)^{\frac{2}{2-\alpha}}] \leq \mathbb{E}[D_{2,\rho,N}(x_1,\ldots,x_N)^2]^{\frac{1}{2-\alpha}}$$

Inserting the bound for the discrepancy obtained in Proposition 4 we end up with the estimate

$$\mathbb{E}\left[\left\|f-f_{\omega}^{\delta}\right\|_{L_{2,\rho}(\Omega)}^{2}\right] \leq \tilde{C}^{\frac{2}{2-\alpha}} N^{-\frac{1}{2-\alpha}} + \frac{2}{2-\alpha} C \delta^{2}$$

For the proposed choices of N we obtain the desired result. $\square$

Remark 13 (Discussion) Observe that the convergence rate in the theorem above is dimension independent. How many points x_i are necessary to obtain a certain quality depends on the noise level δ only; moreover, how fast this value increases in case we obtain measurements with lower noise level does not depend on the space dimensionality; in the worst case we must impose the growth condition $N \sim \delta^{-4}$, under additional smoothness assumptions we only need $N \sim \delta^{-2}$.

On the other hand, the smoothness requirements in the theorems above do depend on the dimension, but as was mentioned above, we already need a certain index of differentiability to ensure that all functions involved are continuous. While of course every classically differentiable function is also continuous, this does not hold for the weakly differentiable functions considered in the definition of Sobolev spaces (see [1, Chap. V] for some counter examples). Therefore stated in terms of Sobolev-norms also the smoothness condition for having a well-defined sampling problem must depend on the dimension.

4 Two Approaches to satisfy Assumptions 1

In this section we demonstrate two possibilities to satisfy Assumptions 1. The particular framework we consider is function approximation with feed forward neural networks with one hidden layer. Although we will now mainly concentrate on one algorithm, observe that the results in Section 3 do in no way depend on the method used to generate the approximations. The results are valid for any method that generates approximations in accordance to Assumptions 1. At the end of this section we briefly demonstrate that also Tikhonov regularization satisfies the required assumptions.

Greedy Approximation

The approximation schemes that we consider for the greedy algorithm have a simple structure, but can still attain high convergence rates; the networks we focus on are given by functions f_k with

$$f_k(x) = \sum_{i=1}^{k} c_i \Phi(x, t_i)\,. \tag{10}$$

The generating function Φ is called *activation function* and could for instance be given by a Gaussian, centered at the point t_i. The main difference between (10) and linear schemes (like e. g., splines) is that the parameters t_i are chosen a-posteriori in dependence of the function f to be approximated. This results in the *dimension independent* convergence rate

$$\|f - f_k\|_H^2 = O\left(\frac{1}{k}\right), \tag{11}$$

where H is some appropriate Hilbert-space (see e. g., [11,12,13,14]). Given point measurements $f^\delta(x_i)$ at points $\omega = \{x_1, \ldots, x_N\}$, the corresponding space is $\ell_2(\omega)$. To associate a function $f_k \in L_2(\Omega)$ with a vector in $\ell_2(\omega)$ (again denoted by f_k) we simply utilize (10) to obtain

$$f_k = \begin{pmatrix} f_{k,1} \\ \vdots \\ f_{k,N} \end{pmatrix} := \begin{pmatrix} f_k(x_1) \\ \vdots \\ f_k(x_N) \end{pmatrix} = \sum_{i=1}^{k} c_i \begin{pmatrix} \Phi(x_1, t_i) \\ \vdots \\ \Phi(x_N, t_i) \end{pmatrix} =: \sum_{i=1}^{k} c_i \begin{pmatrix} \Phi_1(t_i) \\ \vdots \\ \Phi_N(t_i) \end{pmatrix}$$

(implicitly we used the assumption that Φ is continuous, which is of course natural, when we want to consider point measurements in (1)). Using (11), an approximation f_k^δ to f from point values $f^\delta(x_i)$ with $\|f - f_k^\delta\|_{\ell_2(\omega)} \leq \delta$ could in principle be obtained by approximating the vector $f^\delta \in \ell_2(\omega)$ directly with arbitrary high precision, since we have the estimate

$$\left\|f - f_k^\delta\right\|_{\ell_2(\omega)} \leq \left\|f - f^\delta\right\|_{\ell_2(\omega)} + \left\|f^\delta - f_k^\delta\right\|_{\ell_2(\omega)} \leq \delta + O\left(\frac{1}{\sqrt{k}}\right). \tag{12}$$

But as turns out it is not a good idea to approximate f^δ with high accuracy, because with increasing size of the network the generalization properties become worse again: although the error in the $\ell_2(\omega)$-norm decreases, it will not decrease in the $L_2(\Omega)$-norm; the reason for this is that the corresponding approximations will in general not remain smooth, but start to oscillate.

As we have seen in Theorem 8 above, we obtain convergence to f on the whole domain, when the generated approximation f_ω^δ is an approximation of f on $\ell_2(\omega)$ and at the same time remains bounded in a Sobolev space of sufficiently high order. Thus a key to obtain good generalization performance is to maintain boundedness of the norm of f_ω^δ in certain Sobolev spaces. This is in contrast to the scheme in (12), which in general will generate a sequence with $\|f_k^\delta\|_{H^s(\Omega)} \to \infty$ for arbitrary $s > 0$.

Several methods have been proposed for generating smooth approximations; classical attempts to construct such iterates are the Tikhonov-type methods weight decay and output smoothing (investigated in e. g. [15,16]), and the method of early stopping (which has been shown not to be a regularization method in [17]).

Here we want to focus on so-called greedy algorithms, in particular we will use a *weak relaxed greedy algorithm* (see [18] for a recent survey on nonlinear approximation theory): In such an algorithm not all parameters t_i are determined at the same time, but they are obtained one after the other, each one in a locally optimal way. To present the algorithm in more detail we have to introduce some notations first.

We will denote with G_b the set of possible summands ("nodes") in (10), i. e.,

$$G_b = \{ g \in L_2(\Omega) \mid g = c\Phi(\cdot,t), t \in P, |c| \leq b \},$$

where P represents the compact set of attainable parameters. For simplicity we will assume $\|\Phi(\cdot,t)\|_{L_2(\Omega)} \leq 1$. Furthermore we assume that all $g \in G_b$ are continuous, which is vital if we want to consider point-evaluations later.

The greedy algorithm constructs convex approximations and thus generates indices c_i that fulfill the condition $\sum_{i=1}^{k} |c_i| \leq b$; consequently the approximations f_k always lie in the convex hull of G_b, given by

$$\mathrm{co}(G_b) = \left\{ f \in L_2(\Omega) \mid f = \sum_{i=1}^{k} c_i \Phi(\cdot,t_i), \sum_{i=1}^{k} |c_i| \leq b, t_i \in P, k \in \mathbb{N} \right\}.$$

Convergence of f_k to f can of course only be obtained when f is in the closure of this set, which can be written as

$$\overline{\mathrm{co}}(G_b) = \left\{ f \in L_2(\Omega) \mid f = \int_P \Phi(\cdot,t)\, d\mu(t), \|\mu\|_{\mathcal{M}} \leq b \right\},$$

where μ is a Radon-measure and $\|\cdot\|_{\mathcal{M}}$ denotes the corresponding norm. Under the rather natural smoothness assumption $f \in \overline{\mathrm{co}}(G_b)$ we can give the greedy algorithm as shown on the following page.

Note that in the current setting Algorithm 1 requires knowledge of the parameter b; an implementation, which does not require such (in practice unavailable) information has been derived in [19].

The following proposition shows that Algorithm 1 is feasible and generates approximations with a quality in the order of the noise level δ (see [19]).

Proposition 14 *Let $f \in \overline{\mathrm{co}}(G_b)$ and f^δ be such that $\|f - f^\delta\|_{\ell_2(\omega)} \leq \delta$. Then Algorithm 1 is feasible up to the index*

$$k_* = \left\lceil \frac{\eta^2 M_0}{4\delta^2(1+\eta)} \right\rceil. \tag{14}$$

The residual for this index is bounded by

$$\left\| f^\delta - f^\delta_{k_*} \right\|_{\ell_2(\omega)} \leq 2\frac{1+\eta}{\eta}\delta. \tag{15}$$

Algorithm 1. Approximation of noisy data with given smoothness parameter b.

Let $f \in \overline{\text{co}}(G_b)$ and $f^\delta(x_i)$, $i = 1, \dots, N$ as in Assumption 1, then an approximation to f can be computed as follows.

Set $f_0^\delta = 0$.

Choose $M = (1+\eta)M_0$ with M_0 as in (13) and $\eta > 0$.

Compute k_* via (14).

for $k := 1$ **to** k_* **do**

Find $g_k^\delta \in G_b$ such that

$$\left\| f^\delta - \frac{k-1}{k} f_{k-1}^\delta - \frac{1}{k} g_k^\delta \right\|_{\ell_2(\omega)}^2 \leq \frac{M}{k}$$

is fulfilled and define f_k^δ as

$$f_k^\delta = \frac{k-1}{k} f_{k-1}^\delta + \frac{1}{k} g_k^\delta .$$

end for

Here the parameter M_0 is defined as

$$M_0 = \sup_{g \in G_b} \|g\|_{\ell_2(\omega)}^2 - \left\| f^\delta \right\|_{\ell_2(\omega)}^2 + 2\delta \left\| f^\delta \right\|_{\ell_2(\omega)} . \tag{13}$$

PROOF. The statement follows from results in [19], although some caution is necessary to apply these results. In [19], M_0 was defined as $b^2 - \|f^\delta\|^2 + 2\delta\|f^\delta\|$. Here we use the assumption that $f \in \overline{\text{co}}(G_b) \subset L_2(\Omega)$, but consider approximation on $\ell_2(\omega)$. While $\sup_{g \in G_b} \|g\|_{L_2(\Omega)} = b$, the supremum measured in the $\ell_2(\omega)$-norm can be larger. Therefore we explicitly define M_0 as in (13). To check that M_0 is finite, we compute

$$\|g\|_{\ell_2(\omega)}^2 \leq b^2 \sup_{t \in P} \|\Phi(\cdot, t)\|_{\ell_2(\omega)}^2 .$$

Since Φ was assumed to be bounded, we also obtain that M_0 is bounded; the result is now immediately obtained by [19, Theorem 7]. □

The stability of the greedy algorithm stems from the fact that the c_i remain bounded in the l_1-norm throughout the algorithm. Since all f_k^δ have the form $f_k^\delta = \sum_{i=1}^k c_i^k \Phi(\cdot, t_i)$, the restriction $\sum_{i=1}^k \left| c_i^k \right| \leq b$ restricts the class of functions that can be approximated. The chosen bound serves as a regularization parameter with a similar effect as the penalty in Tikhonov-regularization.

Note that the error bound (15) is always larger than 2δ, in contrast to the approach in (12), where the noisy data were approximated with higher accuracy. But although we loose accuracy with respect to the $\ell_2(\omega)$-norm of the error, this solution is preferable since the approximating sequence f_k^δ remains bounded in Sobolev spaces of higher order, in particular it fulfills Assumptions 1 as the next theorem shows.

Theorem 15 *Let the activation function satisfy $\Phi(\cdot,t) \in H^s(\Omega)$ for all $t \in P$, P compact. Then the approximations $f_{k_*}^\delta$ generated by Algorithm 1 fulfill*

- $\|f - f_{k_*}^\delta\|_{\ell_2(\omega)} \leq C\delta$
- $\|f - f_{k_*}^\delta\|_{H^s(\Omega)} \leq \tilde{C}$,

with constants C and $\tilde{C}$ independent of the choice of the points ω. In particular they fulfill the Assumptions 1 with parameter s.

PROOF. The first estimate follows immediately from Proposition 14 via the triangle inequality

$$\begin{aligned}\left\|f - f_{k_*}^\delta\right\|_{\ell_2(\omega)} &\leq \left\|f - f^\delta\right\|_{\ell_2(\omega)} + \left\|f^\delta - f_{k_*}^\delta\right\|_{\ell_2(\omega)} \\ &\leq \delta + 2\frac{1+\eta}{\eta}\delta =: C\delta.\end{aligned}$$

Observe that, although M_0 in (13) does depend on ω, the constant C does not. The second estimate results from the compactness of the set of parameters P (cf. also [20, Sec. 5])

$$\begin{aligned}\left\|f - f_{k_*}^\delta\right\|_{H^s(\Omega)} &\leq \|f\|_{H^s(\Omega)} + \left\|f_{k_*}^\delta\right\|_{H^s(\Omega)} \leq \|f\|_{H^s(\Omega)} + \frac{1}{k}\sum_{i=1}^{k}\left\|g_i^\delta\right\|_{H^s(\Omega)} \\ &\leq 2b\sup_{t\in P}\|\Phi(\cdot,t)\|_{H^s(\Omega)} =: \tilde{C} < \infty \qquad \square\end{aligned}$$

Thus, setting $f_\omega^\delta = f_{k_*}^\delta$ we obtain via Theorem 8 and 12 that $f_{k_*}^\delta$ will converge to f on the whole space $L_2(\Omega)$ for $\delta \to 0$, although Algorithm 1 only uses noisy function values on a discrete, sparse subset $\omega \subset \Omega$.

Tikhonov Regularization

In the following we briefly show that also approximations generated via Tikhonov-regularization satisfy Assumptions 1. Here the corresponding approximation is defined as the minimizer of the functional

$$\left\|f^\delta - f_\alpha^\delta\right\|_{\ell_2(\omega)}^2 + \alpha\left\|f_\alpha^\delta\right\|_{H^s(\Omega)}^2 \to \min_{f_\alpha^\delta \in C \subset H^s(\Omega)} \tag{16}$$

where C is a closed, convex subset of $H^s(\Omega)$ (cf. also to the *regularization networks* treated in [21]). To obtain convergence of this method for noise level δ tending to

0 it is important to choose α properly. A common parameter choice rule is the *discrepancy principle* ([22]).

Remark 16 (Discrepancy Principle) *For given $\tau \geq 1$ and noise level δ as in* (3a)*, choose the largest regularization parameter α for which the minimizer f_α^δ of* (16) *satisfies*

$$\left\| f^\delta - f_\alpha^\delta \right\|_{\ell_2(\omega)} = \tau\delta .$$

We can show that also f_α^δ obtained from Tikhonov regularization together with the discrepancy principle satisfies Assumptions 1.

Theorem 17 *Let $f \in C \subset H^s(\Omega)$, where C is a closed, convex subset of $H^s(\Omega)$ and f_α^δ the minimizer of* (16)*. Furthermore suppose that α is chosen according to the discrepancy principle. Then with $f_\omega^\delta := f_\alpha^\delta$, Assumptions 1 (with parameter s) are satisfied.*

PROOF. Since f_α^δ is a minimizer of (16) we have in particular

$$\left\| f^\delta - f_\alpha^\delta \right\|_{\ell_2(\omega)}^2 + \alpha \left\| f_\alpha^\delta \right\|_{H^s(\Omega)}^2 \leq \left\| f^\delta - f \right\|_{\ell_2(\omega)}^2 + \alpha \| f \|_{H^s(\Omega)}^2 \leq \delta^2 + \alpha \| f \|_{H^s(\Omega)}^2 .$$

Furthermore α was chosen according to the discrepancy principle, this implies

$$\left\| f_\alpha^\delta \right\|_{H^s(\Omega)}^2 \leq \frac{1-\tau^2}{\alpha}\delta^2 + \| f \|_{H^s(\Omega)}^2 \leq \| f \|_{H^s(\Omega)}^2$$

The constants $C_1 = 1 + \tau$ and $C_2 = 2\|f\|_{H^s(\Omega)}$ are independent of ω, therefore $f_\omega^\delta = f_\alpha^\delta$ satisfies Assumptions 1. □

Thus, setting $f_\omega^\delta = f_\alpha^\delta$ we obtain again via Theorem 8 and 12 that $f_{k_*}^\delta$ will converge to f on the whole space $L_2(\Omega)$ for $\delta \to 0$.

5 Outlook

The results presented in this work are valid for sampling on bounded domains. If the domains of interest are unbounded, one can consider transformations that map the unbounded domain onto a bounded one, while at the same time introducing an additional density, which decays towards the boundary. The approach via interpolation inequalities that was presented in Lemma 9 to obtain the improved Theorem 12 used the assumption that $\rho(x) \geq \varepsilon > 0$ and must therefore be replaced by techniques that are also valid for densities that decay to 0. In e. g. [9] and [23] interpolation inequalities for weighted norms have been presented for some particular cases of probability distributions.

Considering unbounded domains it should also be mentioned, that although in the definition of the discrepancy we used a boundary value of the domain as anchor (namely the point $(1,1,\dots,1)$). This does not pose a problem for unbounded domains, since any anchor, i. e., fixed point within the domain may be chosen for defining the discrepancy (see [7]).

In this work we focused on the function approximation problem (1), nevertheless, we would like to mention that also the case of (linear) inverse problems can be treated in an analogous way. Consider the equation

$$Ax = f, \quad f \in \mathcal{R}(A) \subset L_2(\Omega), \tag{17}$$

with some compact operator A, i. e., the range of A, $\mathcal{R}(A)$ is a dense subset of $L_2(\Omega)$. In the classical functional analytic theory of this problem (see e. g., [22]), always full measurements are considered, typically it is assumed that a noisy approximation f^δ is available, where $\|f - f^\delta\|_{L_2(\Omega)} \leq \delta$.

When now discrete measurements are considered, one would again have to verify Assumptions 1. Since a compact operator A is smoothing, f will typically satisfy the smoothness requirements; for standard regularization schemes also f^δ_ω will fulfill (3c). Therefore the results seem to be applicable also for this important class of problems; extension to linear as well as nonlinear inverse problems is a goal of future work.

It should be mentioned that the problem of sparse data in (17) has of course been investigated (see e. g., [24,25,26,27]), but for special cases only. For instance [25] considered the case of Fredholm integral equations; [27] utilized locally polynomial estimators to generate an approximation f^δ_h first, and used Tikhonov-regularization afterwards to solve the inverse problem (17) with right hand side f^δ_h afterwards.

References

[1] R. A. Adams, Sobolev Spaces, Academic Press, 1975.

[2] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Am. Math. Soc., New Ser. 39 (1) (2002) 1–49.

[3] S. Smale, D.-X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Am. Math. Soc., New Ser. 41 (3) (2004) 279–305.

[4] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (1) (2000) 1–50.

[5] P. Niyogi, F. Girosi, Generalization bounds for function approximation from scattered noisy data, Adv. Comput. Math. 10 (1999) 51–80.

[6] H. Niederreiter, Random Number Generation and Quasi-Monte Carlo Methods, SIAM, 1992.

[7] F. J. Hickernell, I. H. Sloan, G. W. Wasilkowski, On tractability of weighted integration over bounded and unbounded regions in $\mathbb{R}^s$, Math. Comp. 73 (2004) 1885–1901.

[8] D. Williams, Probability with martingales, Cambridge University Press, Cambridge, 1991.

[9] R. C. Brown, D. B. Hinton, Weighted interpolation inequalities of sum and product form in $\mathbb{R}^n$, Proc. London Math. Soc. 56 (3) (1988) 261–280.

[10] J. L. Lions, E. Magenes, Non-Homogeneous Boundary Value Problems and Applications, Vol. 1, Springer, Berlin, Heidelberg, 1972.

[11] A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, IEEE Trans. Inform. Theory 39 (3) (1993) 930–945.

[12] A. T. Dingankar, I. W. Sandberg, A note on error bounds for approximation in inner product spaces, Circuits Syst. Signal Process. 15 (4) (1996) 519–522.

[13] M. Donahue, L. Gurvits, C. Darken, E. Sontag, Rates of convex approximation in non-Hilbert spaces, Constructive Approximation 13 (2) (1997) 187–220.

[14] L. K. Jones, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, Ann. Stat. 20 (1) (1992) 608–613.

[15] U. Bodenhofer, M. Burger, H. W. Engl, J. Haslinger, Regularized data-driven construction of fuzzy controllers, J. Inverse Ill-Posed Probl. 10 (2002) 319–344.

[16] M. Burger, A. Neubauer, Error bounds for approximation with neural networks, J. Approx. Theory 112 (2001) 235–250.

[17] A. Hofinger, Iterative regularization and training of neural networks, Diplomarbeit, University of Linz (2003).
URL `http://www.ricam.oeaw.ac.at`

[18] V. Temlyakov, Nonlinear methods of approximation, Found. Comput. Math. 3 (1) (2003) 33–107.

[19] A. Hofinger, Nonlinear function approximation: Computing smooth solutions with an adaptive greedy algorithm, J. Approx. Theory (to appear).

[20] M. Burger, A. Hofinger, Regularized greedy algorithms for network training with data noise, Computing 74 (2005) 1–22.

[21] F. Girosi, M. Jones, T. Poggio, Regularization theory and neural networks architectures, Neural Computation 7 (1995) 219–269.

[22] H. W. Engl, M. Hanke, A. Neubauer, Regularization of Inverse Problems, Kluwer, Dordrecht, 1996.

[23] S.-K. Chua, Weighted Sobolev interpolation inequalities on product spaces., Forum Math. 11 (6) (1999) 647–658.

[24] M. Nashed, G. Wahba, Convergence rates of approximate least squares solutions of linear integral and operator equations of the first kind, Math. Comp. 28 (1974) 69–80.

[25] C. W. Groetsch, C. R. Vogel, Asymptotic theory of filtering for linear operator equations with discrete noisy data, Math. Comp. 49 (1987) 499–506.

[26] M. A. Lukas, Comparison of parameter choice methods for regularization with discrete noisy data, Inverse Probl. 14 (1998) 161–184.

[27] N. Bissantz, T. Hohage, A. Munk, Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise, Inverse Probl. 20 (6) (2004) 1773–1789.