5

Estimates of Model Complexity in Neural-Network Learning

Věra Kůrková

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2, Prague 8, Czech Republic
vera@cs.cas.cz

Abstract. Model complexity in neural-network learning is investigated using tools from nonlinear approximation and integration theory. Estimates of network complexity are obtained from inspection of upper bounds on convergence of minima of error functionals over networks with an increasing number of units to their global minima. The estimates are derived using integral transforms induced by computational units. The role of dimensionality of training data defining error functionals is discussed.

1 Introduction

Many computational models currently used in soft computing can be formally described as devices producing input-output functions in the form of linear combinations of simple computational units corresponding to the model (for example, free-node splines, wavelets, trigonometric polynomials with free frequencies, sigmoidal perceptrons or radial-basis functions). Coefficients of linear combinations as well as inner parameters of computational units are adjustable by various learning algorithms (see, e.g., [1]). Such models have been successfully used in many pattern recognition, optimization, and classification tasks, some of them high-dimensional [2].

In all practical applications, model complexity is constrained. So it is important to choose a type of computational units that allows efficient learning from the given data by networks with a reasonably small number of units. Some insight into impact of the choice of type of units on model complexity can be obtained from investigation of speed of decrease of infima of error functionals over models with an increasing number of computational units. The faster the infima of the error functionals converge to their global minima, the smaller computational models are sufficient for a satisfactory learning from the data determining the functionals.

In this chapter, we derive estimates of rates of convergence of error functionals using tools from approximation and integration theory. Applying upper bounds on rates of nonlinear approximation of neural-network type, we obtain estimates of rates of convergence of error functionals. The estimates are formulated in terms of special norms tailored to various types of computational units. We propose to measure data complexity

M. Bianchini et al. (Eds.): Innovations in Neural Infor. Paradigms & Appli., SCI 247, pp. 97–111.
springerlink.com

with respect to a type of computational units by magnitudes of these norms of functions interpolating the data.

The chapter is organized as follows. In Section 2, basic concepts from learning theory are recalled. In Section 3, some tools from nonlinear approximation theory are presented. In Section 4, certain variational norms tailored to computational units are introduced and a method for their estimation is described. In section 5, upper bounds on rates of convergence of error functionals in terms of variational norms are derived and they are exploited to get estimates of model complexity. In Section 6, the results are illustrated by the example of perceptron networks and the impact of growth of data dimensionality on model complexity is discussed.

2 Learning from Data

Learning from data has been modeled as minimization of *error functionals* over *hypothesis sets* of functions which can be implemented by various computational models (see, e.g., [3], [4]).

Error functionals are determined by training data and loss functions. *Training data* are described either by a discrete sample of input-output pairs or a probability distribution from which such pairs are chosen. A *loss function* measures how much is lost when the model computes $f(x)$ instead of y. The most common loss function is the *quadratic loss* $V(f(x), y) = (f(x) - y)^2$.

A discrete sample $z = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} \,|\, i = 1, \ldots, m\}$ with all x_i distinct ($\mathbb{R}$ denotes the set of real numbers) determines the *empirical error functional* $\mathcal{E}_z$, which is defined for the quadratic loss function as

$$\mathcal{E}_z(f) =: \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2.$$

A nondegenerate (no nonempty open set has measure zero) *probability measure* ρ on the set of input-output pairs $Z = \Omega \times Y$ (where Ω is a *compact* subset of $\mathbb{R}^d$, Y a *bounded* subset of $\mathbb{R}$), determines the *expected error functional* $\mathcal{E}_\rho$ defined for the quadratic loss function as

$$\mathcal{E}_\rho(f) =: \int_Z (f(x) - y)^2 \, d\rho.$$

Hypothesis sets of input-output functions of one-hidden-layer neural networks belong to a class of computational models called *variable-basis* schemes. Such models compute functions from sets of the form

$$\operatorname{span}_n G =: \left\{ \sum_{i=1}^{n} w_i g_i \,\Big|\, w_i \in \mathbb{R},\ g_i \in G \right\},$$

where G is a set of functions, which is sometimes called a *dictionary*. The number n expresses the *model complexity*; in neurocomputing it represents the number of units in the hidden layer.

Typically, dictionaries G are parameterized sets of functions. For example, functions computable by perceptrons, radial-basis-function units, trigonometric polynomials or free-node splines. Such *parameterized families of functions* can be described by mappings

$$\phi : \Omega \times A \to \mathbb{R},$$

where Ω is a set of variables and A is a set of parameters. Usually, $\Omega \subseteq \mathbb{R}^d$ and $A \subseteq \mathbb{R}^p$. We denote by

$$G_\phi = G_\phi(A) =: \{\phi(., a) \,|\, a \in A\}$$

the parameterized set of functions determined by ϕ.

For example, *perceptrons with an activation function* $\sigma : \mathbb{R} \to \mathbb{R}$ induce a mapping ϕ_σ on $\mathbb{R}^{d+1}$ defined for $(v, b) \in \mathbb{R}^d \times \mathbb{R} = \mathbb{R}^{d+1}$ as

$$\phi_\sigma(x, v, b) =: \sigma(v \cdot x + b). \tag{1}$$

Usually, σ is a *sigmoidal function*, i.e., $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to\infty} \sigma(t) = 1$ and σ is nondecreasing. A important sigmoidal is the *Heaviside function* $\vartheta : \mathbb{R} \to \mathbb{R}$ defined as $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$. We denote by $\phi_\vartheta : S^{d-1} \to \mathbb{R}$ (where S^{d-1} denotes the sphere in $\mathbb{R}^d$) the mapping defined as

$$\phi_\vartheta(x, e, b) =: \vartheta(e \cdot x + b). \tag{2}$$

Similarly, *radial-basis functions (RBF)* with a radial function $\beta : \mathbb{R}_+ \to \mathbb{R}$ induce a mapping ϕ_β on $\mathbb{R}^{d+1}$ defined for $(v, b) \in \mathbb{R}^d \times \mathbb{R} = \mathbb{R}^{d+1}$ as

$$\phi_\beta(x, v, b) = \beta(b\|x - v\|). \tag{3}$$

Note that these two types of units are geometrically opposite. Perceptrons compute functions of the form (1), which are plane waves (they are constant on all hyperplanes parallel to the hyperplane $\{x \in \mathbb{R}^d \,|\, v \cdot x + b = 0\}$, and radial units compute functions of the form (3), which are spherical waves (they are constant on spheres centered at v).

Learning algorithms aim to find network parameters generating input-output functions with values of error functionals close to their infima. The algorithms either operate on computational models with a fixed number of units chosen in advance or add units to obtain input-output functions for which values of error functionals better approximate their global minima. In all practical applications, model complexity is constrained and so it is important to choose among potential types of units such types for which error functionals determined by the given training data can achieve smaller values over networks with less units.

Some insight into impact of the choice of computational units on model complexity in learning from a given type of data can be obtained from investigation of *speed of decrease of infima of error functionals* over sets $\mathrm{span}_n G_\phi$ with n increasing. The faster the infima of an error functional converge to its global minimum, the smaller computational model is sufficient for a satisfactory learning from the data determining the functional.

An advantage of the quadratic loss function is that it allows representations of error functionals in terms of distance functionals, for which rates of convergence can be estimated using tools from approximation theory.

The empirical error functional can be expressed as the square of the distance from the function $f_z : \{x_1, \ldots, x_m\} \to \mathbb{R}$ defined as $f(x_i) = y_i$ for all $i = 1, \ldots, m$. The distance is measured in the *weighted l^2-norm* on $\mathbb{R}^m$ defined as

$$\|x\|_{2,m}^2 =: \frac{1}{m} \sum_{i=1}^{m} x_i^2.$$

For $f : \Omega \to \mathbb{R}$, let $f_{|X}$ denote f restricted to $X = \{x_1, \ldots, x_m\}$. Then

$$\mathcal{E}_z(f) = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 = \frac{1}{m} \sum_{i=1}^{m} (f_{|X}(x_i) - f_z(x_i))^2 = \|f_{|X} - f_z\|_{2,m}^2. \quad (4)$$

So the value of the empirical error $\mathcal{E}_z$ at f can be expressed as the square of the l_m^2-distance of f_z from the restriction of f to the set $X = \{x_1, \ldots, x_m\}$.

The expected error functional $\mathcal{E}_\rho$ can be expressed in terms of a distance from the *regression function* f_ρ. This function is defined for all $x \in \Omega$ as

$$f_\rho(x) =: \int_Y y \, d\rho(y|x),$$

where $\rho(y|x)$ is the *conditional (w.r.t. x) probability measure* on Y.

Let $(\mathcal{L}^2(\Omega, \rho_\Omega), \|.\|_{\mathcal{L}^2_{\rho_\Omega}})$ be the space of functions satisfying $\int_\Omega f^2 d\rho_\Omega < \infty$, where ρ_Ω denotes the *marginal probability measure* on Ω defined for every $S \subseteq \Omega$ as $\rho_\Omega(S) = \rho(\pi_\Omega^{-1}(S))$ with $\pi_\Omega : \Omega \times Y \to \Omega$ denoting the projection. The global minimum of the expected error $\mathcal{E}_\rho$ over the whole space $\mathcal{L}^2(\Omega, \rho_\Omega)$ with Ω compact is achieved at f_ρ and for all $f \in \mathcal{L}^2(\Omega, \rho_\Omega)$,

$$\mathcal{E}_\rho(f) = \|f - f_\rho\|_{\mathcal{L}^2_{\rho_\Omega}}^2 + \mathcal{E}_\rho(f_\rho). \quad (5)$$

(see, e.g., [4, p.5]). So $\mathcal{E}_\rho$ can be expressed as the square of the $\mathcal{L}^2_{\rho_\Omega}$-distance from f_ρ plus a constant.

3 Tools from Approximation Theory

Due to the equivalence of minimization of error functionals with the quadratic loss function to minimization of distance functionals, we can study minimization of these functionals using tools from approximation theory. We can take an advantage of Maurey-Jones-Barron's theorem and its corollaries. For functions from the convex hull of a bounded subset G of a Hilbert space, Maurey-Jones-Barron's theorem gives an upper bound on the square of the error in approximation by convex combinations of n elements of G denoted by

$$\mathrm{conv}_n G =: \left\{ \sum_{i=1}^{n} w_i g_i \,\middle|\, w_i \in [0,1],\ \sum_{i=1}^{m} w_i = 1,\ g_i \in G \right\}.$$

The upper bound was originally derived by Maurey (see [5]) using a probabilistic argument. Jones [6] derived a slightly weaker estimate constructively with an iterative

algorithm. Barron [7] refined Jones constructive argument to obtain the same estimate as Maurey. Here, we state their result in a slightly reformulated way with a proof from [8] which is a simplification of Barron's argument. By cl is denoted the *closure* with respect to the topology induced by the ambient space norm.

Theorem 1 (Maurey-Jones-Barron). *Let G be a nonempty bounded subset of a Hilbert space $(\mathcal{X}, \|.\|_{\mathcal{X}})$ and $s_G = \sup_{g\in G} \|g\|_{\mathcal{X}}$, then for every $f \in \mathrm{cl}\,\mathrm{conv}\, G$ and for every positive integer n,*

$$\|f - \mathrm{conv}_n G\|^2_{\mathcal{X}} \leq \frac{s^2_G - \|f\|^2_{\mathcal{X}}}{n}.$$

Proof. Since the distance from $\mathrm{conv}_n G$ is continuous on $\mathcal{X}$ (see e.g., [9]), it is sufficient to verify the statement for $f \in \mathrm{conv}\, G$. Let $f = \sum_{j=1}^m a_j h_j$ be a representation of f as a convex combination of elements of G. Set $c = s^2_G - \|f\|^2_{\mathcal{X}}$.

We show by induction that there exists a sequence $\{g_i\}$ of elements of G such that the barycenters $f_n = \sum_{i=1}^n \frac{g_i}{n}$ satisfy $e^2_n = \|f - f_n\|^2_{\mathcal{X}} \leq \frac{c}{n}$. First check that there exists $g_1 \in G$ such that g_1 satisfies $e^2_1 = \|f - f_1\|^2_{\mathcal{X}} \leq c$. We estimate the convex combination:

$$\sum_{j=1}^m a_j \|f - h_j\|^2_{\mathcal{X}} = \|f\|^2_{\mathcal{X}} - 2f \cdot \sum_{j=1}^m a_j h_j + \sum_{j=1}^m a_j \|h_j\|^2_{\mathcal{X}} \leq s^2_G - \|f\|^2_{\mathcal{X}} = c.$$

Thus there must exist at least one $j \in \{1, \ldots, m\}$ for which $\|f - h_j\|^2_{\mathcal{X}} \leq c$ and we set $f_1 = g_1 = h_j$.

Assuming that we already have $g_1, \ldots, g_n$, we express e^2_{n+1} in terms of e^2_n as

$$e^2_{n+1} = \|f - f_{n+1}\|^2_{\mathcal{X}} = \|\frac{n}{n+1}(f - f_n) + \frac{1}{n+1}(f - g_{n+1})\|^2_{\mathcal{X}} =$$

$$\frac{n^2}{(n+1)^2} e^2_n + \frac{2n}{(n+1)^2}(f - f_n) \cdot (f - g_{n+1}) + \frac{1}{(n+1)^2}\|f - g_{n+1}\|^2_{\mathcal{X}}. \quad (6)$$

Similarly as in the first step, where we considered a convex combination, in this case we also estimate a convex combination of the last two terms from the equation (6):

$$\sum_{j=1}^m a_j \left(\frac{2n}{(n+1)^2}(f - f_n) \cdot (f - h_j) + \frac{1}{(n+1)^2}\|f - h_j\|^2_{\mathcal{X}} \right) =$$

$$\frac{2n}{(n+1)^2}(f - f_n) \cdot (f - \sum_{j=1}^m a_j h_j) + \frac{1}{(n+1)^2} \left(\|f\|^2_{\mathcal{X}} - 2f \cdot (\sum_{j=1}^m a_j h_j) + \sum_{j=1}^m a_j \|h_j\|^2_{\mathcal{X}} \right) =$$

$$\frac{1}{(n+1)^2}(\sum_{j=1}^m a_j g_j - \|f\|^2_{\mathcal{X}}) \leq \frac{1}{(n+1)^2}(s^2_G - \|f\|^2_{\mathcal{X}}) = \frac{c}{(n+1)^2}.$$

Thus there must exist some $j \in \{1, \ldots, m\}$ such that

$$\frac{2n}{(n+1)^2}(f - f_n) \cdot (f - g_{n+1}) + \frac{1}{(n+1)^2}\|f - g_{n+1}\|^2_{\mathcal{X}} \leq \frac{c}{(n+1)^2}.$$

Setting $g_j = h_j$, we get $e^2_{n+1} \leq \frac{n^2}{(n+1)^2} e^2_n + \frac{c}{(n+1)^2}$.

It can be easily verified by induction that this recursive formula together with $e_1^2 \leq c$ gives $e_n^2 \leq \frac{c}{n}$. □

4 Variational Norms

Maurey-Jones-Barron's theorem can be reformulated in terms of a norm called *G-variation*. This norm is defined for any bounded nonempty subset G of any normed linear space $(\mathcal{X}, \|.\|_{\mathcal{X}})$ as the Minkowski functional of the closed convex symmetric hull of G, i.e.,

$$\|f\|_{G,\mathcal{X}} = \|f\|_G =: \inf\left\{c > 0 \mid c^{-1} f \in \mathrm{cl}\,\mathrm{conv}\,(G \cup -G)\right\}, \tag{7}$$

where the closure cl is taken with respect to the topology generated by the norm $\|.\|_{\mathcal{X}}$ and conv denotes the convex hull. Note that G-variation can be infinite. It is a norm on the subspace of $\mathcal{X}$ formed by those $f \in \mathcal{X}$, for which $\|f\|_G < \infty$. G-variation depends on the norm on the ambient space, but when this is implicit, we omit it in the notation.

Variational norms were introduced by Barron [10] for characteristic functions of certain families of subsets of $\mathbb{R}^d$, in particular, for the set of characteristic functions of closed half-spaces of the form $\{x \in \mathbb{R}^d \mid e \cdot x + b \geq 0\}$, which correspond to the set of functions computable by Heaviside perceptrons. For functions of one variable (i.e., $d = 1$), the *variation with respect to half-spaces* coincides, up to a constant, with the notion of total variation [10,11]. The general concept was defined in [12]. The next corollary from [12] (see also [8]) gives an upper bound on rates of approximation by $\mathrm{span}_n G$ for all functions in a Hilbert space.

Corollary 1. *Let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a Hilbert space, G its bounded nonempty subset, $s_G = \sup_{g \in G} \|g\|_{\mathcal{X}}$. Then for every $f \in \mathcal{X}$ and every positive integer n,*

$$\|f - \mathrm{span}_n G\|_{\mathcal{X}}^2 \leq \frac{s_G^2 \|f\|_G^2 - \|f\|_{\mathcal{X}}^2}{n}.$$

Proof. Let $b = \|f\|_G$. If b is infinite, the statement holds trivially, otherwise set $G' = bG = \{b\,g \mid g \in G\}$. By Theorem 1, $\|f - \mathrm{conv}_n G'\|_{\mathcal{X}} \leq \frac{s_{G'}^2 - \|f\|_{\mathcal{X}}^2}{n}$. As $\mathrm{conv}_n G' \subseteq \mathrm{span}_n G$, we have $\|f - \mathrm{span}_n G\|_{\mathcal{X}} \leq \|f - \mathrm{conv}_n G'\|_{\mathcal{X}} \leq \frac{(s_G\, b)^2 - \|f\|_{\mathcal{X}}^2}{n} = \frac{s_G^2 \|f\|_G^2 - \|f\|_{\mathcal{X}}^2}{n}$. □

To apply Maurey-Jones-Barron's theorem to neural networks, one has to estimate variational norms with respect to common types of network units. One method of such estimation exploits integral representations of functions in the form of "networks with infinitely many hidden units" which instead of finite linear combinations $\sum_{i=1}^n w_i \phi(., a_i)$ compute

$$\int_A w(a)\phi(x,a)d\mu(a). \tag{8}$$

Many functions can be expressed as such networks with infinitely many units or as limits of sequences of such networks. For example, all smooth functions, which are either compactly supported or sufficiently rapidly decreasing to zero, can be expressed

as networks with infinitely many perceptrons [11,13]. All continuous compactly supported functions and all $\mathcal{L}^p$-functions with $p \in [1, \infty)$ are limits of convolutions with suitable kernels including the Gaussian one and thus they are limits of networks with kernel and radial units [14,15] and all smooth functions (functions in Sobolev spaces) can be expressed as networks with infinitely many Gaussian radial units [16,17].

The following theorem from [18] shows that for functions representable as infinite networks of the form (8), G_ϕ-variation can be estimated by the $\mathcal{L}^1$-norm of the output weight function w. For the Lebesgue measure λ, we write shortly $\mathcal{L}^p(\Omega) = \mathcal{L}^p_\lambda(\Omega)$.

Theorem 2. *Let $\Omega \subseteq \mathbb{R}^d$ be Lebesgue measurable, $f \in \mathcal{L}^p(\Omega)$, $p \in [1, \infty)$, be such that for all $x \in \Omega$,*

$$f(x) = \int_A w(a)\phi(x, a)da,$$

where $\phi : \Omega \times A \to \mathbb{R}$ is such that $G_\phi(A) = \{\phi(., a) \,|\, a \in A\}$ is a bounded subset of $(\mathcal{L}^p(\Omega), \|.\|_{\mathcal{L}^p})$ and $w \in \mathcal{L}^1(A)$. Then

$$\|f\|_{G_\phi(A)} \leq \|w\|_{\mathcal{L}^1(A)}.$$

Note that various special cases of this theorem were proven earlier: the case of ϕ being a trigonometric function [7] and a general theorem requiring compactness of the parameter set A or continuity of the hidden unit function ϕ [11,19]. Theorem 2 has minimal assumptions which are necessary for formulation of the estimate $\|f\|_{G_\phi(A)} \leq \|w\|_{\mathcal{L}^1(A)}$. The set $G_\phi(A)$ has to be bounded so that $G_\phi(A)$-variation is defined and the output-weight function w has to be in $\mathcal{L}^1(A)$ so that the upper bound is defined.

5 Data Complexity with Respect to Computational Units

Often, neither the regression function f_ρ nor any function interpolating the sample z is computable by a network of a given type. Even if some of these functions can be represented as an input-output function of a network from the given class, the network might have too many hidden units to be implementable.

For all common types of computational units, the sets

$$\operatorname{span} G_\phi = \cup_{n=1}^\infty \operatorname{span}_n G_\phi$$

are dense in $\mathcal{L}^p$-spaces and in the space $(\mathcal{C}(\Omega), \|.\|_{\sup})$ of continuous functions with the supremum norm with $\Omega \subset \mathbb{R}^d$ compact (see, e.g., [20], [21] and the references therein). The next proposition show that the global minima of error functionals over $\mathcal{L}^2$-spaces are equal to their infima over sets of functions computable by neural networks.

Proposition 1. *(i) Let both $\Omega \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ be compact, ρ be a nondegenerate probability measure on $Z = \Omega \times Y$, f_ρ the regression function, and $\phi : \Omega \times Y \to \mathbb{R}$ be such that* $\operatorname{span} G_\phi$ *is dense in $\mathcal{L}^2_{\rho_\Omega}(\Omega)$. Then*

$$\mathcal{E}_\rho(f_\rho) = \min_{f \in \mathcal{L}^2_{\rho_\Omega}(\Omega)} \mathcal{E}_\rho(f) = \inf_{f \in \operatorname{span} G_\phi} \mathcal{E}_\rho(f) = \lim_{n \to \infty} \inf_{f \in \operatorname{span}_n G_\phi} \mathcal{E}_\rho(f);$$

(ii) Let $\Omega \subseteq \mathbb{R}^d$, $z = \{(x_i, y_i) \in \Omega \times \mathbb{R} \,|\, i = 1, \ldots, m\}$ with all x_i distinct, and $\phi : \Omega \times Y \to \mathbb{R}$ be such that $\operatorname{span} G_\phi$ is dense in $(\mathcal{C}(\Omega), \|.\|_{\sup})$. Then

$$0 = \min_{f \in \mathcal{L}^2_{\mu\Omega}} \mathcal{E}_z(f) = \inf_{f \in \operatorname{span} G_\phi} \mathcal{E}_z(f) = \lim_{n \to \infty} \inf_{f \in \operatorname{span}_n G_\phi} \mathcal{E}_z(f).$$

Proof. The representations (5) and (4), resp., show that $\mathcal{E}_\rho$ is continuous on $\mathcal{L}^2_{\rho_\Omega}(\Omega)$ and $\mathcal{E}_z$ is continuous on $\mathcal{C}(\Omega)$. It is easy to see that a minimum of a continuous functional over the whole space is equal to its infimum over any dense subset. □

So the infima of error functionals over sets $\operatorname{span}_n G_\phi$ converge with n increasing to their global minima. Note that when sets of hidden unit functions G_ϕ are linearly independent, then sets $\operatorname{span}_n G_\phi$ are not convex and thus results from theory of convex optimization cannot be applied. So we have to consider merely infima over sets $\operatorname{span}_n G_\phi$ because minima might not be achieved.

The *speed of convergence* of these infima with n increasing to the global minima is critical for capability of networks with reasonable numbers of units computing ϕ to learn from the data described by the distribution ρ or the sample z. Inspection of estimates of this speed can suggest some characterizations of *complexity of the data* with respect to the given type of computational units.

The following theorem show that G_ϕ-variation can play a role of a measure of complexity of data with respect to the computational units computing the function ϕ.

Theorem 3. *(i) Let both $\Omega \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ be compact, ρ be a nondegenerate probability measure on $Z = \Omega \times Y$, f_ρ the regression function, and G be a bounded subset of $\mathcal{L}^2(\Omega, \rho_\Omega)$ with $s_G = \sup_{g \in G} \|g\|_{\mathcal{L}^2_{\rho_\Omega}}$. Then for all n*

$$\inf_{f \in \operatorname{span}_n G} \mathcal{E}_\rho(f) - \mathcal{E}_\rho(f_\rho) \leq \frac{s_G^2 \|f_\rho\|_G^2}{n};$$

(ii) Let d, m be positive integers, $\Omega \subseteq \mathbb{R}^d$, μ a measure on Ω, $z = \{(x_i, y_i) \in \Omega \times \mathbb{R} \,|\, i = 1, \ldots, m\}$ with all x_i distinct, and G be a bounded subset of $\mathcal{L}^2(\Omega, \mu)$ with $s_G = \sup_{g \in G} \|g\|_{\mathcal{L}^2_\mu(\Omega)}$. Then for every $h \in \mathcal{L}^2_\mu(\Omega)$ interpolating the sample z,

$$\inf_{f \in \operatorname{span}_n G} \mathcal{E}_z(f) \leq \frac{s_G^2 \|h\|_G^2}{n}.$$

Proof. By the representation (5), for every $f \in \mathcal{L}^2_{\rho_\Omega}(\Omega)$, $\mathcal{E}_\rho(f) - \mathcal{E}_\rho(f_\rho) = \|f_\rho - f\|^2_{\mathcal{L}^2_{\rho_\Omega}}$ and so $\inf_{f \in \operatorname{span}_n G} \mathcal{E}_\rho(f) - \mathcal{E}_\rho(f_\rho) = \|f_\rho - \operatorname{span}_n G\|^2_{\mathcal{L}^2_{\rho_\Omega}}$. Thus it remains to estimate the distance of f_ρ from $\operatorname{span}_n G$. By Corollary 1, this distance is bounded from above by $\frac{s_G \|f_\rho\|_G}{\sqrt{n}}$. So $\inf_{f \in \operatorname{span}_n G} \mathcal{E}_\rho(f) - \mathcal{E}_\rho(f_\rho) \leq \frac{s_G^2 \|f_\rho\|_G^2}{n}$.

Let $G_{|X}$ denote the set of functions from G restricted to $X = \{x_1, \ldots, x_m\}$. By the representation (4), for every $f \in \mathcal{L}^2_\mu(\Omega)$, $\mathcal{E}_z(f) = \|f_{|X} - f_z\|^2_{2,m}$, where $f_{|X}$ denotes f restricted to X. So $\inf_{f \in \operatorname{span}_n G} \mathcal{E}_z(f) = \|f_z - \operatorname{span}_n G_{|X}\|^2_{2,m}$. By Corollary 1, $\|f_z - \operatorname{span}_n G_{|X}\|_{2,m} \leq \frac{s_{G_{|X}} \|f_z\|_{G_{|X}}}{\sqrt{n}}$. Hence $\inf_{f \in \operatorname{span}_n G} \mathcal{E}_z(f) \leq \frac{s^2_{G_{|X}} \|f_z\|^2_{G_{|X}}}{n}$. It follows

directly from the definition of variational norm that if $h_{|X} = f_z$, then $\|f_z\|_{G_{|X}} \leq \|h\|_G$. Thus for every h interpolating the sample z,
$\inf_{f \in \mathrm{span}_n G} \mathcal{E}_z(f) \leq \frac{s_G^2 \|h\|_G^2}{n}$. □

Theorem 3 implies estimates of the number of network units sufficient for a given accuracy of approximation of the global minima of error functionals.

Corollary 2. *(i) Let both $\Omega \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ be compact, ρ be a non degenerate probability measure on $Z = \Omega \times Y$, f_ρ the regression function, and $G_\phi = \{\phi(.,y) \,|y \in Y\}$ be a bounded subset of $\mathcal{L}^2(\Omega, \rho_\Omega)$ with $s_\phi = \sup_{y \in Y} \|\phi(.,y)\|_{\mathcal{L}^2_{\rho_\Omega}}$, and $\varepsilon > 0$. Then for all $n \geq \frac{s_\phi^2 \|f_\rho\|^2_{G_\phi}}{\varepsilon}$, the infimum of the functional $\mathcal{E}_\rho$ over a network with n units computing ϕ is within ε from its global minimum $\mathcal{E}_\rho(f_\rho)$ over the whole space $\mathcal{L}^2(\Omega, \rho_\Omega)$;*
(ii) Let d, m be positive integers, $\Omega \subseteq \mathbb{R}^d$, μ a measure on Ω, and $z = \{(x_i, y_i) \in \Omega \times \mathbb{R} \,|\, i = 1, \ldots, m\}$ with all x_i distinct, G_ϕ be a bounded subset of $\mathcal{L}^2(\Omega, \mu)$ with $s_G = \sup_{g \in G} \|g\|_{\mathcal{L}^2_\mu(\Omega)}$, and $\varepsilon > 0$. Then for every n such that for some function $h \in \mathcal{L}^2_\mu(\Omega)$ which interpolates the sample z,

$$n \geq \frac{s_\phi^2 \|h\|^2_{G_\phi}}{\varepsilon},$$

holds, the infimum of $\mathcal{E}_z$ over a network with n units computing ϕ is smaller or equal to ε.

So the infima of error functionals achievable over sets $\mathrm{span}_n G_\phi$ decrease at least as fast as $\frac{1}{n}$ times the square of the G_ϕ-variational norm of the regression function or some interpolating function. When these norms are small, good approximations of the global minima of error functionals can be obtained using networks with a moderate number of units.

The critical factor in these estimates is the magnitude of the G_ϕ-variational norm of the function from which the training data are chosen (the regression function f_ρ or any function interpolating the discrete sample). Thus comparing magnitudes of G_ϕ-variations for various types of computational unit functions ϕ, one can get some understanding how model complexity of a neural network is influenced by the choice of a type of its units. The magnitudes of the G_ϕ-variational norms of the regression function or some function interpolating the sample z can be used as measures of complexity of data given by the probability measure ρ or a finite sample z with respect to units computing ϕ.

6 Data Complexity with Respect to Perceptrons

To apply Corollary 2 to perceptrons, we need to estimate variation with respect to perceptrons. As for all sigmoidals σ, G_{ϕ_σ}-variation in $\mathcal{L}^2(\Omega)$ with Ω compact is equal to G_{ϕ_ϑ}-variation [11], it is sufficient to estimate variation with respect to Heaviside perceptrons. This norm is sometimes called *variation with respect to half-spaces* [10] as

perceptons with the Heaviside activation function compute characteristic functions of closed half-spaces of $\mathbb{R}^d$ intersected with Ω.

Variation with respect to half-spaces of smooth functions can be estimated applying Theorem 2 to a representation of such functions as networks with infinitely many Heaviside perceptrons. The following theorem gives such a representation for all functions from $\mathcal{C}^d(\mathbb{R}^d)$ (functions having all partial derivatives up to the order d continuous), which are of a *weakly controlled decay*. These are the functions which satisfy for all multi-indexes α with $0 \leq |\alpha| = \alpha_1 + \ldots \alpha_d < d$, $\lim_{\|x\|\to\infty}(D^\alpha f)(x) = 0$ (where $D^\alpha = (\partial/\partial x_1)^{\alpha_1} \ldots (\partial/\partial x_d)^{\alpha_d}$) and for some $\varepsilon > 0$, all multi-indexes α with $|\alpha| = d$ satisfy

$$\lim_{\|x\|\to\infty}(D^\alpha f)(x)\|x\|^{d+1+\varepsilon} = 0.$$

Note that the class of functions of weakly controlled decay contains the class of all compactly supported functions from $\mathcal{C}^d(\mathbb{R}^d)$) and the Schwartz class $\mathcal{S}(\mathbb{R}^d)$ (all functions from $\mathcal{C}^\infty(\mathbb{R}^d)$ which are together with all their derivatives rapidly decreasing [22, p. 251]). In particular, the Gaussian function $\gamma_d(x) = \exp(-\|x\|^2)$ belongs to the class of functions of a weakly controlled decay.

By S^{d-1} is denoted the *unit sphere in* $\mathbb{R}^d$, by $D_e^{(d)}$ the *directional derivative of the order* d *in the direction* e, and by $H_{e,b} = \{x \in \mathbb{R}^d \,|\, x \cdot e + b = 0\}$ the *hyperplane* determined by the normal vector $e \in S^{d-1}$ and the bias b.

Theorem 4. *Let* d *be an odd integer and* $f \in \mathcal{C}^d(\mathbb{R}^d)$ *be of a weakly controlled decay, then for all* $x \in \mathbb{R}^d$

$$f(x) = \int_{S^{d-1}\times\mathbb{R}} w_f(e,b)\,\vartheta(e\cdot x + b)\,de\,db,$$

where $w_f(e,b) = a(d)\int_{H_{e,b}} D_e^{(d)}(f)(y)\,dy$ *and* $a(d) = (-1)^{(d-1)/2}(1/2)(2\pi)^{1-d}$.

Theorem 5 shows that many smooth functions can be expressed as networks with infinitely many Heaviside perceptrons. The output-weight functions $w_f(e,b)$ in such a network is the product of a function $a(d)$ of the number of variables d converging with d increasing exponentially fast to zero and a "flow of order d through the hyperplane" $H_{e,b}$.

The integral representation from Theorem 5 was first derived by Ito [23] for all functions from the Schwartz class. Ito used the Radon transform (see, e.g., [22, p.251]) to prove universal approximation property of perceptron networks and the representation from Theorem 5 is not explicitly stated in the paper [23], but can be obtained by combining Theorem 3.1, Proposition 2.2 and an equation on p.387 in [23]. In [11] the same formula was derived for all compactly supported functions from $\mathcal{C}^d(\mathbb{R}^d)$, d odd, via an integral formula for the Dirac delta function. In [24], the integral representation from Theorem 5 was extended to functions of weakly controlled decay. Note that in Theorem 5, we have stated the integral representation of smooth functions as networks with infinitely many Heaviside perceptrons only for d odd, but a similar representation also holds for d even, but the output weight function is much more complicated

All proofs of above mentioned representations require rather advanced tools. Here we present the proof for compactly supported functions in $\mathcal{C}^d(\mathbb{R}^d)$ from

[11, Theorem 4.1], which is relatively self-contained. The proof is takes an advantage of the representation of the Heaviside function as the first distributional derivative of the Dirac delta function and a representation of the d-dimensional *Dirac delta function* δ_d as an integral of one-dimensional Dirac delta functions δ_1. We use the following relationship between d-dimensional and one-dimensional Dirac delta functions from [25, p.680].

Proposition 2. *For every odd positive integer d*

$$\delta_d(x) = a_d \int_{S^{d-1}} {\delta_1}^{(d-1)}(e \cdot x)\, de,$$

where $a_d = (-1)^{\frac{d-1}{2}}/(2(2\pi)^{d-1})$.

We first prove a technical lemma.

Lemma 1. *For all positive integers d,k, every $f \in \mathcal{C}^d(\mathbb{R}^d)$, every $e \in S^{d-1}$, and every $b \in \mathbb{R}$, $\frac{\partial^k}{\partial b^k} \int_{H_{e,b}} f(y)\, dy = \int_{H_{e,b}} {D_e}^{(k)} f(y)\, dy$.*

Proof. First, we verify that the statement is true for $k = 1$:
$\frac{\partial}{\partial b} \int_{H_{e,b}} f(y)\, dy = \lim_{t\to 0} t^{-1} \left(\int_{H_{\mathbf{e}b}} f(y) dy - \int_{H_{eb+t}} f(y)\, dy \right) =$
$\lim_{t\to 0} t^{-1} \int_{H_{e,b}} (f(y + te) - f(y))\, dy = \int_{H_{e,b}} \lim_{t\to 0} t^{-1} (f(y + te) - f(y))\, dy =$
$\int_{H_{e,b}} D_e f(y)\, dy$.

Suppose that the statement is true for $k - 1$. Then
$\frac{\partial^k}{\partial b^k} \int_{H_{e,b}} f(y)\, dy = \lim_{t\to 0} t^{-1} \left(\int_{H_{\mathbf{e}b}} D_e^{(k-1)} f(y)\, dy - \int_{H_{eb+t}} D_e^{(k-1)} f(y)\, dy \right) =$
$\lim_{t\to 0} t^{-1} \int_{H_{e,b}} \left(D_e^{(k-1)} f(y + te) - D_e^{(k-1)} f(y) \right) dy =$
$\int_{H_{e,b}} \lim_{t\to 0} t^{-1} \left(D_e^{(k-1)} f(y + te) - D_e^{(k-1)} f(y) \right) dy = \int_{H_{e,b}} D_e^{(k)} f(y)\, dy.$ □

Proof of Theorem 5. We first prove the theorem for test functions, i.e., compactly supported functions from $\mathcal{C}^d(\mathbb{R}^d)$. For any test function f by the definition of the delta distribution we have $f(x) = (f * \delta_d)(x) = \int_{\mathbb{R}^d} f(z)\delta_d(x - z)\, dz$ (see e.g., [26]). By Proposition 2, $\delta_d(x - z) = a_d \int_{S^{d-1}} {\delta_1}^{(d-1)}(e \cdot x - e \cdot z)\, de$. Thus, $f(x) = a_d \int_{S^{d-1}} \int_{\mathbb{R}^d} f(z)\, {\delta_1}^{(d-1)}(x \cdot e - z \cdot e)\, dz\, de$. So rearranging the inner integration, we get

$$f(x) = a_d \int_{S^{d-1}} \int_{\mathbb{R}} \int_{H_{e,b}} f(y)\, {\delta_1}^{(d-1)}(x \cdot e + b)\, dy\, db\, de.$$

Setting $u(e, b) = a_d \int_{H_{e,b}} f(y)\, dy$, we obtain
$f(x) = \int_{S^{d-1}} \int_{\mathbb{R}} u(e, b) {\delta_1}^{(d-1)}(x \cdot e + b)\, db\, de$.

By the definition of distributional derivative, $\int_{\mathbb{R}} u(e, b) {\delta_1}^{(d-1)}(e \cdot x + b)\, db =$
$(-1)^{d-1} \int_{\mathbb{R}} \frac{\partial^{d-1} u(e,b)}{\partial b^{d-1}} \delta_1(e \cdot x + b)\, db$ for every $e \in S^{d-1}$ and $x \in \mathbb{R}^d$. Since d is odd, we have $\int_{\mathbb{R}} u(e, b) {\delta_1}^{(d-1)}(e \cdot x + b)\, db = \int_{\mathbb{R}} \frac{\partial^{d-1} u(e,b)}{\partial b^{d-1}} \delta_1(e \cdot x + b)\, db$.

Since the first distributional derivative of the Heaviside function is the delta distribution (see e.g.,[26, p. 47]), it follows that for all $e \in S^{d-1}$ and $x \in \mathbb{R}^d$ $\int_{\mathbb{R}} u(e,b)\delta_1{}^{(d-1)}(e\cdot x+b)db = -\int_{\mathbb{R}} \frac{\partial^d u(e,b)}{\partial b^d}\vartheta(e\cdot x+b)db$.

By Lemma 1, $\frac{\partial^d u(e,b)}{\partial b^d} = \frac{\partial^d}{\partial b^d}\int_{H_{e,b}} f(y)\,dy = \int_{H_{e,b}} D_e^{(d)} f(y)\,dy$. Hence,

$$f(x) = -a_d \int_{S^{d-1}} \int_{\mathbb{R}} \left(\int_{H_{e,b}} D_{\mathbf{e}}^{(d)} f(y) dy \right) \vartheta(e\cdot x+b)\,db\,de.$$

Now let $f \in \mathcal{C}^d(\mathbb{R}^d)$ be compactly supported. Then there exists a sequence $\{f_i\}$ of test functions converging to f uniformly on $\mathbb{R}^d$ (see e.g., [26, p.3]. It is easy to check that for every $e \in S^{d-1}$, $\{D_e^{(d)} f_i\}$ converges uniformly on $\mathbb{R}^d$ to $D_e^{(d)} f$. Hence we can interchange limit and integration (see e.g., [27, p.233]) to obtain $\lim_{i\to\infty}\int_{H_{e,b}} D_e^{(d)} f_i(y)\,dy = \int_{H_{e,b}} D_e^{(d)} f(y)\,dy$. Let $g_i(x,e,b) =$
$\int_{H_{e,b}} \left(D_e^{(d)} f_i(y)\,dy\right)\vartheta(e\cdot x+b)$ and $g(x,e,b) = \int_{H_{e,b}} \left(D_{\mathbf{e}}^{(d)} f(y)dy\right)\vartheta(e\cdot x+b)$. Then it is easy to see that for all $x\in\mathbb{R}^d$, $\lim_{i\to\infty} g_i(x,e,b) = g(x,e,b)$ uniformly on $S^{d-1}\times\mathbb{R}$. Thus for all $x\in\mathbb{R}^d$,

$$f(x) = \lim_{i\to\infty}\int_{S^{d-1}}\int_{\mathbb{R}} g_i(x,e,b)\,db\,de = \int_{S^{d-1}}\int_{\mathbb{R}} g(x,e,b)dbde =$$

$$\int_{S^{d-1}}\int_{\mathbb{R}} \left(\int_{H_{eb}} D_e^{(d)} f(y)dy)\right)\vartheta(e\cdot x+b)\,db\,de$$

(using again interchangebility of integration and limit for a sequence of functions converging uniformly). □

Combining Theorems 2 and 3, we obtain the following corollary.

Corollary 3. *Let d be an odd integer and $f\in\mathcal{C}^d(\mathbb{R}^d)$ be of a weakly controlled decay, then*

$$\|f\|_{G_{\phi_\vartheta}(\Omega),\mathcal{L}^2} \le \|f\|_{G_{\phi_\vartheta}(\mathbb{R}^d),\sup} \le \|w_f\|_{\mathcal{L}^1(\mathbb{R}^d)},$$

where $w_f(e,b) = a(d)\int_{H_{e,b}} D_e^{(d)}(f)(y)\,dy$ and $a(d) = (-1)^{(d-1)/2}(1/2)(2\pi)^{1-d}$.

The $\mathcal{L}^1$-norm of the weighting function w_f can be estimated by a product of a function

$$k(d) \sim \left(\frac{4\pi}{d}\right)^{1/2}\left(\frac{e}{2\pi}\right)^{d/2} < \left(\frac{4\pi}{d}\right)^{1/2}\left(\frac{1}{2}\right)^{d/2}$$

with the Sobolev seminorm $\|f\|_{d,1,\infty}$ of the function f [24]. The *seminorm* $\|.\|_{d,1,\infty}$ is defined as

$$\|f\|_{d,1,\infty} = \max_{|\alpha|=d}\|D^\alpha f\|_{\mathcal{L}^1(\mathbb{R}^d)},$$

where $\alpha = (\alpha_1,\ldots,\alpha_d)$ is a multi-index with nonnegative integer components, $D^\alpha = (\partial/\partial x_1)^{\alpha_1}\ldots(\partial/\partial x_d)^{\alpha_d}$ and $|\alpha| = \alpha_1+\cdots+\alpha_d$.

Thus by Corollary 3

$$\|f\|_{G_{\phi_\vartheta}(\mathbb{R}^d),\sup} \le \|w_f\|_{\mathcal{L}^1(\mathbb{R}^d)} \le k(d)\|f\|_{d,1,\infty} = k(d) \max_{|\alpha|=d} \|D^\alpha f\|_{\mathcal{L}^1(\mathbb{R}^d)} \tag{9}$$

where $k(d)$, which is *decreasing exponentially fast with the number of variables* d.

Note that for large d, the seminorm $\|f\|_{1,d,\infty}$ is much smaller than the standard Sobolev norm $\|f\|_{d,1} = \sum_{|\alpha|\le d} \|D^\alpha f\|_{\mathcal{L}^1(\mathbb{R}^d)}$ [22] as instead of the *summation of* 2^d *iterated partial derivatives* of f over all α with $|\alpha| \le d$, merely their *maximum* over α with $|\alpha| = d$ is taken.

The following theorem estimates speed of decrease of minima of error functionals over networks with an increasing number n of Heaviside perceptrons.

Theorem 5. *Let d, m be positive integers, d odd, both $\Omega \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ be compact, $z = \{(x_i, y_i) \in \Omega \times Y \mid i = 1, \ldots, m\}$ with all x_i distinct, ρ be a non degenerate probability measure on $\Omega \times Y$, such that the regression function $f_\rho : \Omega \to \mathbb{R}$ is a restriction of a function $h_\rho \in \mathcal{C}^d(\mathbb{R}^d)$ of a weakly controlled decay and let $h \in \mathcal{C}^d(\mathbb{R}^d)$ be a function of a weakly controlled decay interpolating the sample z. Then for all n*

$$\min_{f \in \mathrm{span}_n G_{\phi_\vartheta}(\Omega)} \mathcal{E}_z(f) \le \frac{c(d)\|h\|^2_{d,1,\infty}}{n}$$

$$\text{and} \quad \min_{f \in \mathrm{span}_n G_{\phi_\vartheta}(\Omega)} \mathcal{E}_\rho(f) - \mathcal{E}_\rho(f_\rho) \le \frac{c(d)\|h_\rho\|^2_{d,1,\infty}}{n},$$

where $c(d) \sim \frac{4\pi}{d}\left(\frac{e}{2\pi}\right)^d < \frac{4\pi}{d2^d}$.

Proof. It was shown in [28] that each function in $\mathcal{L}^2_{\rho_\Omega}(\Omega)$ has its best approximations in sets $\mathrm{span}_n G_{\phi_\vartheta}$ for all n. Thus by (5) and (4), both the functionals $\mathcal{E}_\rho$ and $\mathcal{E}_z$ achieve over $\mathrm{span}_n G_{\phi_\vartheta(\Omega)}$ their minima. By (9), for all d odd and all h of a weakly controlled decay

$$\|h\|_{G_{\phi_\vartheta}(\Omega),\mathcal{L}^2} \le k(d)\|h\|_{d,1,\infty},$$

where $k(d) \sim \left(\frac{4\pi}{d}\right)^{1/2}\left(\frac{e}{2\pi}\right)^{d/2}$. The statement follows by Theorem 3. □

Theorem 5 shows that when a sample of data z can be interpolated by a function $h \in \mathcal{C}^d(\mathbb{R}^d)$ which is vanishing sufficiently quickly at infinity and the squares of the maxima of the $\mathcal{L}^1$-norms of its partial derivatives of the order $|\alpha| = d$ do not exceed an exponentially increasing upper bound $\frac{d}{4\pi}2^d$, i. e.,

$$\|h\|^2_{d,1,\infty} = \max_{|\alpha|=d} \|D^\alpha f\|^2_{\mathcal{L}^1_\lambda(\mathbb{R}^d)} \le \frac{1}{c(d)} \sim \frac{d}{4\pi}\left(\frac{2\pi}{e}\right)^d < \frac{d}{4\pi}2^d, \tag{10}$$

then the minima of the empirical error $\mathcal{E}_z$ defined by the sample z over networks with n sigmoidal perceptrons decrease to zero rather quickly – at least as fast as $\frac{1}{n}$.

The estimate (10) shows that with increasing dimensionality of data, tolerance on its "oscillatory behavior" measured by the partial derivatives of an interpolating function

h is increasing exponentially fast. For example, for $d > 4\pi$, when all the $\mathcal{L}^1$-norms of the partial derivatives of the order d of h are close to 2^d, a convergence faster than $\frac{1}{n}$ is guaranteed.

Our estimates of data complexity can be illustrated by the example of the Gaussian function $\gamma_d(x) = \exp(-\|x\|^2)$. It was shown in [24] that for d odd, $\|\gamma_d\|_{G_{\phi_\vartheta}(\Omega)} \leq 2d$ (see also [29] for a weaker estimate depending on the size of Ω, which is valid also for d even). Thus by Theorem 3, when the regression function $f_\rho = \gamma_d$ and the sample z of the size m is such that the $\gamma_{d\,|X} = f_z$, then

$$\min_{f \in \mathrm{span}_n G_{\phi_\vartheta}(\Omega)} \mathcal{E}_\rho(f) \leq \frac{4d^2}{n} \quad \text{and} \quad \min_{f \in \mathrm{span}_n G_{\phi_\vartheta}(\Omega)} \mathcal{E}_z(f) \leq \frac{4d^2}{n}.$$

This estimate gives some insight into a relationship between two geometrically opposite types of computational units - *Gaussian radial-basis functions* and *Heaviside perceptrons*. Minima of the error functionals defined by data chosen from the d-dimensional Gaussian over networks with n Heaviside perceptrons converge to zero faster than $\frac{4d^2}{n}$. Thus to approximate their global minima within ε, it is sufficient to use a network with $n \geq \frac{4d^2}{\varepsilon}$ units. Note that the upper bound $\frac{4d^2}{\varepsilon}$ grows with the dimension d only quadratically and it does not depend on the size m of a sample.

On the other hand, there exist samples $z = \{(x_i, y_i) \,|\, i = 1, \ldots, m\}$, the sizes of which influence the magnitudes of the variations of the functions f_z defined as $f_z(x_i) = y_i$. For example, for any positive integer k, consider $\Omega = [0, 2k]$, $Y = [-1, 1]$ and the sample $z = \{(2i, 1), (2i+1, -1) \,|\, i = 0, \ldots, k-1\}$ of the size $m = 2k$. Then one can easily verify that $\|f_z\|_{G_{\phi_\vartheta}} = 2k$ (for functions of one variable, variation with respect to half-spaces is up to a constant equal to their total variation, see [10], [11]). This example indicates that the more the data "oscillate", the larger the variation with respect to half-spaces of functions interpolating such data.

Acknowledgement

This work was partially supported by the Ministry of Education of the Czech Republic, project Center of Applied Cybernetics 1M684077004 (1M0567) and by the Institutional Research Plan AV0Z10300504.

References

1. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn. Prentice-Hall, Englewood Cliffs (1998)
2. Kecman, V.: Learning and Soft Computing. MIT Press, Cambridge (2001)
3. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)
4. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bulletin of AMS 39, 1–49 (2002)
5. Pisier, G.: Remarques sur un résultat non publié de B. Maurey. In: Séminaire d'Analyse Fonctionnelle 1980-1981, École Polytechnique, Centre de Mathématiques, Palaiseau, France, vol. I(12) (1981)

6. Jones, L.K.: A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. Annals of Statistics 20, 608–613 (1992)
7. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory 39, 930–945 (1993)
8. Kůrková, V.: High-dimensional approximation and optimization by neural networks. In: Suykens, J., Horváth, G., Basu, S., Micchelli, C., Vandewalle, J. (eds.) Advances in Learning Theory: Methods, Models and Applications, ch. 4, pp. 69–88. IOS Press, Amsterdam (2003)
9. Rudin, W.: Best approximation in normed linear spaces by elements of subspaces. Springer, Berlin (1970)
10. Barron, A.R.: Neural net approximation. In: Narendra, K. (ed.) Proc. 7th Yale Workshop on Adaptive and Learning Systems. Yale University Press (1992)
11. Kůrková, V., Kainen, P.C., Kreinovich, V.: Estimates of the number of hidden units and variation with respect to half-spaces. Neural Networks 10, 1061–1068 (1997)
12. Kůrková, V.: Dimension-independent rates of approximation by neural networks. In: Warwick, K., Kárný, M. (eds.) Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality, pp. 261–270. Birkhäuser, Basel (1997)
13. Kainen, P.C., Kůrková, V., Vogt, A.: Integral combinations of heavisides. In: Mathematische Nachrichten (to appear, 2009)
14. Park, J., Sandberg, I.: Universal approximation using radial–basis–function networks. Neural Computation 3, 246–257 (1991)
15. Park, J., Sandberg, I.: Approximation and radial basis function networks. Neural Computation 5, 305–316 (1993)
16. Girosi, F.: Approximation error bounds that use VC- bounds. In: Proceedings of the International Conference on Artificial Neural Networks, Paris, pp. 295–302 (1995)
17. Kainen, P.C., Kůrková, V., Sanguineti, M.: Complexity of Gaussian radial basis networks approximating smooth functions. J. of Complexity 25, 63–74 (2009)
18. Kůrková, V.: Model complexity of neural networks and integral transforms. In: Polycarpou, M., Panayiotou, C., Alippi, C., Ellinas, G. (eds.) Artificial Neural Networks - ICANN 2009. LNCS, vol. 5768, pp. 708–718. Springer, Heidelberg (2009)
19. Kainen, P.C., Kůrková, V.: An integral upper bound for neural network approximation. Neural Computation 21(10), 2970–2989 (2009)
20. Pinkus, A.: Approximation theory of the MLP model in neural networks. Acta Numerica 8, 277–283 (1998)
21. Kůrková, V.: Neural networks as universal approximators. In: Arbib, M. (ed.) The Handbook of Brain Theory and Neural Networks, pp. 1180–1183. MIT Press, Cambridge (2002)
22. Adams, R.A., Fournier, J.J.F.: Sobolev Spaces. Academic Press, Amsterdam (2003)
23. Ito, Y.: Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory. Neural Networks 4, 385–394 (1991)
24. Kainen, P.C., Kůrková, V., Vogt, A.: A Sobolev-type upper bound for rates of approximation by linear combinations of Heaviside plane waves. J. of Approximation Theory 147, 1–10 (2007)
25. Courant, R., Hilbert, D.: Methods of Mathematical Physics, vol. II. Interscience, New York (1994)
26. Zemanian, A.H.: Distribution Theory and Transform Analysis. Dover, New York (1987)
27. Edwards, C.H.: Advanced Calculus of Several Variables. Dover, New York (1994)
28. Kainen, P.C., Kůrková, V., Vogt, A.: Best approximation by Heaviside perceptron networks. Neural Networks 13, 695–697 (2000)
29. Cheang, G.H.L., Barron, A.R.: A better approximation for balls. IEEE Transactions on Information Theory 104, 183–203 (2000)