Estimates of data 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. Complexity of data with respect to a particular class of neural networks is studied. Data complexity is measured by the magnitude of a certain norm of either the regression function induced by a probability measure describing the data or a function interpolating a sample of input/output pairs of training data chosen with respect to this probability. The norm is tailored to a type of computational units in the network class. It is shown that for data for which this norm is "small", convergence of infima of error functionals over networks with increasing number of hidden units to the global minima is relatively fast. Thus for such data, networks with a reasonable model complexity can achieve good performance during learning. For perceptron networks, the relationship between data complexity, data dimensionality and smoothness is investigated.

1 Introduction

The goal of a supervised learning is to adjust parameters of a neural network so that it approximates with a sufficient accuracy a functional relationship between inputs and outputs known only by a sample of *empirical data* (input-output pairs). Many learning algorithms (such as the back-propagation [21], [6]) iteratively decrease the average square of errors on a training set. Theoretically, such learning is modeled as minimization of error functionals defined by data: the *expected error* is determined by data in the form of a *probability measure* and the *empirical error* by a discrete *sample of data* chosen with respect to this measure (see, e.g., [20], [5]).

In most learning algorithms, either the number of network computational units is chosen in advance or it is dynamically allocated, but in both cases, it is constrained. The speed of decrease of infima of error functionals over networks with increasing number of computational units can play a role of a measure of complexity of data with respect to a given type of computational units (such as perceptrons with a given activation function or radial or kernel units with a given kernel).

In this paper, we investigate *data complexity with respect to a class of networks* for data defining the error functionals: a probability measure ρ and a sample of input-output pairs $z = \{(u_i, v_i) \,|\, i = 1, \ldots, m\}$. We derive an upper

bound on the speed of decrease of infima of error functionals over networks with n hidden units depending on a certain *norm tailored to the type of hidden units* of either the regression function defined by the probability measure describing the data or its discrete approximation in the form of a function interpolating a sample of input-output pairs of training data. We show that the speed of decrease od these infima is bounded from above by $\frac{1}{n}$ times the square of this norm. Thus over a network with the number of hidden units n greater than $\frac{1}{\varepsilon}$ times the square of this norm, infima of error functionals are within ε from their global minima. We propose to characterize data complexity by the magnitudes of this norm of the regression or an interpolating function.

For perceptron networks, we investigate the relationship between data complexity, smoothness and dimensionality. We estimate the norm tailored to perceptrons by the product of a function $k(d)$ of the dimension of the data d (which is decreasing exponentially fast to zero) and a Sobolev seminorm of the regression or an interpolating function defined as the maximum of the $\mathcal{L}^1$-norms of the partial derivatives of the order d. This estimate shows that for perceptron networks with increasing dimensionality of inputs, the tolerance on smoothness of the training data (measured by the Sobolev seminorm of the regression or an interpolatig function), which allow learning by networks of a reasonable size, is increasing exponentially fast.

The paper is organized as follows. In section 2, learning is described as minimization of error functionals expressed in terms of distance functionals. In section 3, tools from approximation theory are applied to obtain upper bounds on rates of decrease of infima of the error functionals over networks with increasing model complexity and by inspection of these bounds, a measure of data complexity is proposed. In section 4, the proposed concept of data complexity is illustrated by the example of the class of perceptron networks, for which the relationship between data complexity, data dimensionality and smoothness of the regression or an interpolating function is analyzed.

2 Learning as minimization of error functionals

Let ρ be a non degenerate (no nonempty open set has measure zero) *probability measure* defined on $Z = X \times Y$, where X is a *compact* subset of $\mathbb{R}^d$ and Y a *bounded* subset of $\mathbb{R}$ ($\mathbb{R}$ denotes the set of real numbers). The measure ρ induces the *marginal probability measure* on X defined for every $S \subseteq X$ as $\rho_X(S) = \rho(\pi_X^{-1}(S))$, where $\pi_X : X \times Y \to X$ denotes the projection. Let $(\mathcal{L}^2_{\rho_X}(X), \|.\|_{\mathcal{L}^2_{\rho_X}})$ denote the Lebesque space of functions satisfying $\int_X f^2 d\rho_X < \infty$. The *expected error functional* $\mathcal{E}_\rho$ determined by ρ is defined for every f in $\mathcal{L}^2_{\rho_X}(X)$ as

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

and the *empirical error functional* $\mathcal{E}_z$ determined by a *sample of data* $z = \{(u_i, v_i) \in X \times Y \,|\, i = 1, \ldots, m\}$ is defined as

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

It is easy to see and well-known [5] that the expected error $\mathcal{E}_\rho$ achieves its minimum over the whole space $\mathcal{L}^2_{\rho_X}(X)$ at the *regression function* f_ρ defined for all $x \in X$ 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. Thsus

$$\min_{f \in \mathcal{L}^2_{\rho_X}(X)} \mathcal{E}_\rho(f) = \mathcal{E}_\rho(f_\rho).$$

Moreover,

$$\mathcal{E}_\rho(f) = \int_X (f(x) - f_\rho(x))^2 d\rho_X + \mathcal{E}_\rho(f_\rho) = \|f - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} + \mathcal{E}_\rho(f_\rho)$$

[5, p.5]. So $\mathcal{E}_\rho$ can be expressed as the square of the $\mathcal{L}^2_{\rho_X}$-distance from f_ρ plus a constant

$$\mathcal{E}_\rho(f) = \|f - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} + \mathcal{E}_\rho(f_\rho). \tag{1}$$

The empirical error $\mathcal{E}_z$ achieves its minimum over the whole space $\mathcal{L}^2_{\rho_X}(X)$ at any function that interpolates the sample z, i.e., at any function $h \in \mathcal{L}^2_{\rho_X}(X)$ such that $h_{|X_u} = h_z$, where $X_u = \{u_1, \ldots, u_m\}$ and $h_z : X_u \to Y$ is defined as

$$h_z(u_i) = v_i. \tag{2}$$

For all such functions h,

$$\min_{f \in \mathcal{L}^2_{\rho_X}(X)} \mathcal{E}_z(f) = \mathcal{E}_z(h).$$

Also the empirical error can be expressed in terms of a distance functional. For any $X \subset \mathbb{R}^d$ containing X_u and $f : X \to \mathbb{R}$, let

$$f_u = f_{|X_u} : X_u \to \mathbb{R}$$

denote f restricted to X_u and $\|.\|_{2,m}$ denote the *weighted ℓ_2-norm* on $\mathbb{R}^m$ defined by $\|x\|^2_{2,m} = \frac{1}{m} \sum_{i=1}^m x_i^2$. Then

$$\mathcal{E}_z(f) = \frac{1}{m} \sum_{i=1}^{m} (f(u_i) - v_i)^2 = \frac{1}{m} \sum_{i=1}^{m} (f_u(u_i) - h_z(u_i))^2 = \|f_u - h_z\|^2_{2,m} = \mathcal{E}_z(f_u).$$

So the empirical error $\mathcal{E}_z$ can be expressed as the square of the l^2_m-distance from h_z

$$\mathcal{E}_z(f) = \|f_u - h_z\|^2_{2,m}. \tag{3}$$

3 Characterization of data complexity with respect to a class of networks

To model neural-network learning, one has to consider minimization of error functionals over subsets of $\mathcal{L}^2_{\rho_X}(X)$ formed by functions computable by various classes of networks. 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 class, the network might have too many hidden units to be implementable. In most learning algorithms, either the number of hidden units is chosen in advance or it is dynamically allocated, but in both cases, it is constrained. We investigate complexity of the data ρ and z defining the error functionals with respect to a given class of networks in terms of model complexity of networks sufficient for learning from these data.

The most common class of *networks with n hidden units and one linear output unit* can compute functions of the form

$$span_n\, G = \left\{ \sum_{i=1}^{n} w_i g_i \,|\, w_i \in \mathbb{R},\, g_i \in G \right\},$$

where G is the set of functions that can be computed by *computational units* of a given type (such as perceptrons or radial-basis functions). The number n of hidden units plays the role of a measure of *model complexity of the network*. Its size is critical for a feasibility of an implementation.

For all common types of computational units, the union $\cup_{n=1}^{\infty} span_n G$ of the nested family of sets of functions computable by nets with n hidden units is dense in $\mathcal{L}^2_{\rho_X}(X)$ (see, e.g., [18], [13] and the references therein).

Both the expected and the empirical error functionals are continuous on $\mathcal{L}^2_{\rho_X}(X)$ (their representations (1) and (3) show that they can be expressed as squares of the $\mathcal{L}^2_{\rho_X}$-norm or weighted ℓ^2-norm resp., plus a constant). 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. Thus

$$\inf_{f \in \cup_{n=1}^{\infty} span_n G} \mathcal{E}_\rho(f) = \mathcal{E}_\rho(f_\rho) \qquad \text{and} \qquad \inf_{f \in \cup_{n=1}^{\infty} span_n G} \mathcal{E}_z(f) = 0.$$

Note that for G linearly independent, sets $span_n G$ are not convex and thus results from theory of convex optimization cannot be applied. Thus we have to consider merely $\inf_{f \in span_n G} \mathcal{E}_\rho(f)$ because for a general set G, minima over sets $span_n G$ might not be achieved.

The *speed of convergence* with the number of hidden units n increasing of the infima of error functionals over sets $span_n G$ to the global minima over the whole space $\mathcal{L}^2_{\rho_X}(X)$ is critical for learning capability of the class of networks with hidden units computing functions from G (for example, perceptrons with a certain activation function). Inspection of estimates of this speed can suggest some characterization of *complexity of data* guaranteeing a possibility of learning from such data by networks with a reasonable number of hidden units computing

functions from the class G. We shall show that one such characterization of complexity of data with respect to a class of networks is the *magnitude of a norm tailored to the type of hidden units* of either the regression function f_ρ or any function h interpolating the sample z, i.e., a function satisfying $h(u_i) = v_i$ for all $i, \ldots, m$. If the magnitude of such norm is "small", infima of error functionals over $span_n G$ converge quickly.

The norm, called G-variation, can be defined for any bounded nonempty subset G of a normed linear space $(X, \|.\|)$ (here, we consider the Hilbert space $\mathcal{L}^2_{\rho_X}(X)$ and some parameterized sets G corresponding to sets of functions computable by neural networks). G-variation is defined as the *Minkowski functional of the closed convex symmetric hull of* G, i.e.,

$$\|f\|_G = \inf\left\{c > 0 : \; c^{-1} f \in cl\, conv\,(G \cup -G)\right\}, \tag{4}$$

where the closure cl is taken with respect to the topology generated by the norm $\|.\|$ and $conv$ denotes the convex hull. Note that G-variation can be infinite (when the set on the right-hand side is empty). It was defined in [12] as an extension of the *variation with respect to half-spaces* introduced for Heaviside perceptron networks in [2] (for the properties of variation see [14]).

The following theorem estimates speed of convergence of the infima of the expected and the empirical error functionals over sets $span_n G$ formed by functions computable by networks with n hidden units computing functions from G.

Theorem 1. *Let d, m, n be positive integers, both $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ be compact, $z = \{(u_i, v_i) \in X \times Y \,|\, i = 1, \ldots, m\}$ with all u_i distinct, ρ be a non degenerate probability measure on $X \times Y$, and G be a bounded subset of $\mathcal{L}^2_{\rho_X}(X)$ with $s_G = \sup_{g \in G} \|g\|_{\mathcal{L}^2_{\rho_X}}$. Then*

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

and for every $h \in \mathcal{L}^2_{\rho_X}(X)$ interpolating the sample z,

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

Proof. By the representation (1), for every $f \in \mathcal{L}^2_{\rho_X}(X)$, $\mathcal{E}_\rho(f) - \mathcal{E}_\rho(f_\rho) = \|f_\rho - f\|^2_{\mathcal{L}^2_{\rho_X}}$ and so $\inf_{f \in span_n G} \mathcal{E}_z(f) - \mathcal{E}_\rho(f_\rho) = \|f_\rho - span_n G\|^2_{\mathcal{L}^2_{\rho_X}}$. Thus it remains to estimate the distance of f_ρ from $span_n G$. By an estimate of rates of approximation by $span_n G$ in a Hilbert space derived by Maurey [19], Jones [8] and Barron [2, 3], and reformulated in terms of G-variation in [14], this distance is bounded from above by $\frac{s_G \|f_\rho\|_G}{\sqrt{n}}$. Hence $\inf_{f \in span_n G} \mathcal{E}_\rho(f) - \mathcal{E}_\rho(f_\rho) \leq \frac{s_G^2 \|f_\rho\|_G^2}{n}$.

Let $G_{|X_u}$ denote the set of functions from G restricted to $X_u = \{u_1, \ldots, u_m\}$. By the representation (3), for every $f \in \mathcal{L}^2_{\rho_X}(X)$, $\mathcal{E}_z(f) = \|f_u - h_z\|^2_{\mathcal{L}^2_{\rho_X}}$ and so

$\inf_{f \in span_n G} \mathcal{E}_z(f) = \|h_z - span_n G_{|X_u}\|^2_{2,m}$. By Maurey-Jones-Barron's estimate, $\|h_z - span_n G_{|X_u}\|_{2,m} \leq \frac{s_{G_{|X_u}} \|h_z\|_{G_{|X_u}}}{\sqrt{n}}$. Hence $\inf_{f \in span_n G} \mathcal{E}_z(f) \leq \frac{s^2_{G_{|X_u}} \|h_z\|^2_{G_{|X_u}}}{n}$. It follows directly from the definitions that if $f_{|X_u} = f_u$, then $\|f_u\|_{G_{|X_u}} \leq \|f\|_G$. Thus for every h interpolating the sample z,
$\inf_{f \in span_n G} \mathcal{E}_z(f) \leq \frac{s^2_G \|h\|^2_G}{n}$. □

So the infima of error functionals achievable over networks with n hidden units computing functions from a set G decrease at least as fast as $\frac{1}{n}$ times the square of the G-variational norm of the regression function or some interpolating function. When these norms are small, good approximations of the two global minima, $\min_{f \in \mathcal{L}^2_{\rho_X}(X)} \mathcal{E}_\rho(f) = \mathcal{E}_\rho(f_\rho)$ and $\min_{f \in \mathcal{L}^2_{\rho_X}(X)} \mathcal{E}_z(f) = 0$, can be obtained using networks with a moderate number of units. Thus the magnitudes of the G-variational norms of the regression function or some function interpolating the sample z of input-output pairs can be used as measures of *complexity of data* given by the probability measure ρ or a finite sample z chosen from $X \times Y$ with respect to ρ. When these magnitudes are "small", data have a reasonable complexity for learning by networks with hidden units computing functions from the set G.

4 Smoothness and data complexity with respect to perceptron networks

To get some insight into complexity of data with respect to various types of networks, one has to estimate corresponding variational norms. One method of such estimation takes an advantage of integral representations of functions in the form of "networks with continua of hidden units".

Typically, sets G describing computational units are of the form

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

where $\phi : X \times A \to \mathbb{R}$.

For example, *perceptrons* compute functions from the set

$$P_d(\psi, X) = \{f : X \to \mathbb{R} \mid f(x) = \psi(v_i \cdot x + b_i), v_i \in \mathbb{R}^d, b_i \in \mathbb{R}\},$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is an *activation function* (typically, a *sigmoidal*, i.e., a monotonic nondecreasing function $\sigma : \mathbb{R} \to \mathbb{R}$ satisfying $\lim_{t \to -\infty} \sigma(t) = 0$ and $\lim_{t \to \infty} \sigma(t) = 1$). An important type of a sigmoidal is the *Heaviside function* $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$. So, $P_d(\psi, X) = \{\phi(x, (v_1, \ldots, v_d, b)) \,|\, v \in \mathbb{R}^d, b \in \mathbb{R}\}$, where $\phi(x, (v_1, \ldots, v_d, b)) = \psi(v \cdot x + b)$.

If for X and A compact, a continuous function $f : X \to \mathbb{R}$ can be represented as a "neural network" with a continuum of hidden units computing functions $\phi(., a)$ and with output weights $w(a)$, i.e.,

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

and the weighing function w is in $\mathcal{L}^1_\lambda(X)$, where λ denotes the Lebesgue measure, then

$$\|f\|_G \leq \|w\|_{\mathcal{L}^1_\lambda} \tag{5}$$

[15, Theorem 3.1] (see also [7] and [11] for extensions of this result). So G-variational norm can be estimated using the $\mathcal{L}^1_\lambda$-norm of the weighting function.

For standard computational units, many functions can be represented as such "infinite" networks and moreover the $\mathcal{L}^1_\lambda$-norms of weighting functions can be estimated in terms of some norms expressing certain kinds of smoothness, the upper bound (5) gives a method for estimating the data complexity proposed in the previous section.

For all sigmoidals σ, $P_d(\sigma, X)$-variation in $\mathcal{L}^2_{\rho_X}(X)$ is equal to $P_d(\vartheta, X)$-variation [15]. Thus to investigate complexity with respect to sigmoidal perceptron networks, it is sufficient to estimate variation with respect to Heaviside perceptrons called *variation with respect to half-spaces* (perceptons with the Heaviside activation compute characteristic functions of half-spaces of $\mathbb{R}^d$ intersected with X). To simplify notation, we write

$$H_d(X)$$

instead of $P_d(\vartheta, X)$. So $\|.\|_{H_d} = \|.\|_{P_d(\sigma)}$ for all sigmoidals.

An integral representation as a network with Heaviside perceptrons holds for functions from a wide class (including functions on $\mathbb{R}^d$, which are compactly supported or merely "rapidly decreasing at infinity" and have continuous partial derivatives of all orders) [15], [10]. For d odd, the representation is of the form

$$f(x) = \int_{S^{d-1}\times\mathbb{R}} w_f(e,b)\vartheta(e\cdot x + b)dedb, \tag{6}$$

where S^{d-1} denotes the unit sphere in $\mathbb{R}^d$ and the weighing function $w_f(e,b)$ is a product of a function $a(d)$ of the number of variables d converging with d increasing exponentially fast to zero and a "flow of the order d through the hyperplane" $H_{e,b} = \{x \in \mathbb{R}^d \,|\, x\cdot e + b = 0\}$. More precisely,

$$w_f(e,b) = a(d)\int_{H_{e,b}} (D^{(d)}_e(f))(y)dy,$$

where

$$a(d) = (-1)^{(d-1)/2}(1/2)(2\pi)^{1-d}$$

and $D^{(d)}_e$ denotes the directional derivative of the order d in the direction e.

The integral representation (6) was derived in [15] for compactly supported functions from $\mathcal{C}^d(\mathbb{R}^d)$ and extended in [11] to functions of a *weakly controlled decay*, which satisfy for all α with $0 \leq |\alpha| < d$, $\lim_{\|x\|\to\infty}(D^\alpha f)(x) = 0$ and there exists $\varepsilon > 0$ such that for each multi-index α with $|\alpha| = d$, $\lim_{\|x\|\to\infty}(D^\alpha f)(x)\|x\|^{d+1+\varepsilon} = 0$. The class of functions with weakly controlled decay contains all d-times continuously differentiable functions with compact support as well as all functions

from the Schwartz class $\mathcal{S}(\mathbb{R}^d)$ [1, p.251]). In particular, it contains the Gaussian function $\gamma_d(x) = \exp(-\|x\|^2)$.

In [10], the $\mathcal{L}^1_\lambda$-norm of the weighting function w_f was estimated by a product of a function $k(d)$, which is *decreasing exponentially fast* with the number of variables d, with the Sobolev seminorm of the represented function f:

$$\|w_f\|_{\mathcal{L}^1_\lambda} \le k(d)\|f\|_{d,1,\infty}.$$

The *seminorm* $\|.\|_{d,1,\infty}$ is defined as

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

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 (5)

$$\|f\|_{H_d} \le k(d)\|f\|_{d,1,\infty} = k(d) \max_{|\alpha|=d} \|D^\alpha f\|_{\mathcal{L}^1_\lambda(\mathbb{R}^d)} \tag{7}$$

where

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

Note that for large d, the seminorm $\|f\|_{1,d,\infty}$ is much smaller than the standard Sobolev norm $\|f\|_{d,1} = \sum_{|\alpha|\le d} \|D^\alpha f\|_{\mathcal{L}^1_\lambda(\mathbb{R}^d)}$ [1] 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 increasing number n of Heaviside perceptrons.

Theorem 2. *Let d, m, n be positive integers, d odd, both $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ be compact, $z = \{(u_i, v_i) \in X \times Y \mid i = 1, \ldots, m\}$ with all u_i distinct, ρ be a non degenerate probability measure on $X \times Y$, such that the regression function $f_\rho : X \to \mathbb{R}$ is a restriction of a function $h_\rho : \mathbb{R}^d \to \mathbb{R}$ of a weakly controlled decay and let $h : \mathbb{R}^d \to \mathbb{R}$ be a function of a weakly controlled decay interpolating the sample z. Then*

$$\min_{f \in span_n H_d(X)} \mathcal{E}_z(f) \le \frac{c(d)\|h\|^2_{d,1,\infty}}{n}$$

$$\text{and} \quad \min_{f \in span_n H_d(X)} \mathcal{E}_\rho(f) - \mathcal{E}_\rho(f_\rho) \le \frac{c(d)\|h_\rho\|^2_{d,1,\infty}}{n},$$

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 [9] that sets $span_n H_d(X)$ are approximatively compact in $\mathcal{L}^2_{\rho_X}(X)$ and so each function in $\mathcal{L}^2_{\rho_X}(X)$ has its best approximation in sets $span_n H_d$. Thus by (1) and (3), both the functionals $\mathcal{E}_\rho$ and $\mathcal{E}_z$ achieve over

$span_n H_d$ their minima. It follows from [10] (Theorems 3.3, 4.2 and Corollary 3.4) that for all d odd and all h of a weakly controlled decay

$$\|h\|_{H_d(X)} \leq k(d)\|h\|_{d,1,\infty},$$

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 1. □

Thus for any sample of data z, which can be interpolated by a function $h \in \mathcal{C}^d(\mathbb{R}^d)$ vanishing sufficiently quickly at infinity such that the squares of the maxima of the $\mathcal{L}^1_\lambda$-norms of partial derivatives of the order $|\alpha| = d$ do not exceed an exponentially increasing upper bound $\frac{d}{4\pi}2^d$, more precisely

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

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

For example when for $d > 4\pi$, all the $\mathcal{L}^1_\lambda$-norms of the partial derivatives of the order d are smaller than 2^d, 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 [10] that for d odd, $\|\gamma_d\|_{H_d} \leq 2d$ (see also [4] for a weaker estimate depending on the size of X, which is valid also for d even). Thus by Theorem 1, when the regression function $f_\rho = \gamma_d$ and the sample z of the size m is such that the function h_z defined as $h_z(u_i) = v_i$ is the restriction of the Gaussian function γ_d to $X_u = \{u_1, \ldots, u_m\}$, then

$$\min_{f \in span_n H_d(X)} \mathcal{E}_\rho(f) \leq \frac{4d^2}{n} \text{ and } \min_{f \in span_n H_d(X)} \mathcal{E}_z(f) \leq \frac{4d^2}{n}. \tag{8}$$

This estimate gives some insight into the relationship between two geometrically opposite types of computational units - *Gaussian radial-basis functions* (RBFs) and *Heaviside perceptrons*. Perceptrons compute *plane waves* (functions of the form $\psi(v \cdot x + b)$, which are constant on the hyperplanes parallel with the hyperplane $\{x \in \mathbb{R}^d \,|\, v \cdot x + b = 0\}$), while Gaussian RBFs compute *radial waves* (functions of the from $exp(-(b\|x-v\|)^2)$, which are constant on spheres centered at v). By (8) minima of the error functionals defined by the d-dimensional Gaussian probability measure over networks with n Heaviside perceptrons converge to zero faster than $\frac{4d^2}{n}$. Note that the upper bound $\frac{4d^2}{n}$ 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 $z = \{(u_i, v_i) \,|\, i = 1, \ldots, m\}$, the sizes of which influence the magnitudes of the variations of the functions h_z defined as $h_z(u_i) = v_i$. For example, for any positive integer k, consider $X = [0, 2k]$, $Y = [-1, 1]$ 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 $\|h_z\|_{H_d(X)} = 2k$ (for functions of one variable, variation with respect to half-spaces is up to a constant equal to their total variation, see [2], [15]). This example indicates that the more the data "oscillate", the larger the variation of functions, which interpolate them.

5 Discussion

We proposed a measure of data complexity with respect to a class of neural networks based on inspection of an estimate of speed of convergence of the error functionals defined by the data. For data with a "small" complexity expressed in terms of a magnitude of a certain norm (which is tailored to the network type) of the regression or an interpolating function defined by the data, networks with a reasonable model complexity can achieve good performance during learning.

Our analysis of data complexity in neural-network learning merely considers minimization of error functionals. The next step should be to extend the study to the case of regularized expected errors as in the case of kernel models in [16], [17]. Various stabilizers could be considered, among which variation with respect to half-spaces seems to be the most promising. In one dimensional case, variation with respect to half-spaces is up to a constant equal to total variation [2], [15], which is used as a stabilizer in image processing. Moreover, our estimates show its importance in characterization of data complexity in learning by perceptron networks.

Acknowledgement

This work was partially supported by the project 1ET100300419 "Intelligent Models, Algorithms, Methods, and Tools for Semantic Web Realization" of the National Research Program of the Czech Republic and the Institutional Research Plan AV0Z10300504.

References

1. Adams, R. A., Fournier, J. J. F.: Sobolev Spaces. Academic Press, Amsterdam, 2003.
2. Barron, A. R.: Neural net approximation. Proc. 7th Yale Workshop on Adaptive and Learning Systems, K. Narendra, Ed., pp. 69–72. Yale University Press, 1992.
3. Barron, A. R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory **39** (1993) 930–945.
4. Cheang, G. H. L., Barron, A. R.: A better approximation for balls. Journal of Approximation Theory **104** (2000) 183–200.
5. Cucker, F. , Smale, S.: On the mathematical foundations of learning. Bulletin of AMS **39** (2002) 1–49.
6. Fine, T. L.: Feedforward Neural Networks Methodology. Springer, New York, 1999.
7. Girosi, F., Anzellotti, G.: Rates of convergence for radial basis functions and neural networks. Artificial Neural Networks for Speech and Vision, R. J. Mammone, Ed., pp. 97-113. Chapman & Hall, London, 1993.
8. 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** (1992) 608–613.
9. Kainen, P. C., Kůrková, V., Vogt, A.: Best approximation by linear combinations of characteristic functions of half-spaces. Journal of Approximation Theory **122** (2003) 151–159.

10. Kainen, P. C., Kůrková, V., Vogt, A.: A Sobolev-type upper bound for rates of approximation by linear combinations of plane waves. Submitted, Research report ICS–900, www.cs.cas.cz/research/publications.shtml.
11. Kainen, P. C., Kůrková, V., Vogt, A.: Integral combinations of Heavisides. Submitted, Research report ICS–966, www.cs.cas.cz/research/publications.shtml.
12. Kůrková, V.: Dimension–independent rates of approximation by neural networks. Computer–Intensive Methods in Control and Signal Processing: Curse of Dimensionality, K. Warwick and M. Kárný, Eds., pp. 261–270. Birkhäuser, Boston, 1997.
13. Kůrková, V.: Neural networks as universal approximators. The Handbook of Brain Theory and Neural Networks II, M. Arbib, Ed., pp. 1180–1183. MIT Press, Cambridge, 2002.
14. Kůrková, V.: High-dimensional approximation and optimization by neural networks. Advances in Learning Theory: Methods, Models and Applications(Chapter 4), J. Suykens et al., Eds., pp. 69–88. IOS Press, Amsterdam, 2003.
15. Kůrková, V., Kainen, P. C., V. Kreinovich, V. : Estimates of the number of hidden units and variation with respect to half-spaces. Neural Networks **10** (1997) 1061–1068.
16. Kůrková, V., M. Sanguineti, M.: Error estimates for approximate optimization by the extended Ritz method. SIAM Journal on Optimization **15** (2005) 461–487.
17. Kůrková, V., Sanguineti, M. Learning with generalization capability by kernel methods of bounded complexity. Journal of Complexity **21** (2005) 350–367.
18. Pinkus, A.: Approximation theory of the MPL model in neural networks. Acta Numerica **8** (1998) 277–283.
19. Pisier, G.: Remarques sur un résultat non publié de B. Maurey. Séminaire d'Analyse Fonctionnelle 1980-81, Exposé no. V, pp. V.1-V.12, École Polytechnique, Centre de Mathématiques, Palaiseau, France, 1980.
20. Vapnik, V. N. : The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995.
21. Werbos, P. J.: Backpropagation: Basics and new developments. The Handbook of Brain Theory and Neural Networks. Arbib M., Ed., pp. 134–139, MIT Press, Cambridge, 1985.